

# Fluid Models of Parallel Service Systems under FCFS

Yuval Nov\*

Gideon Weiss†

Hanqin Zhang‡

April 18, 2016

## Abstract

We study deterministic fluid approximations of parallel service systems operating under first come first served policy (FCFS). The condition for complete resource pooling is identified in terms of the system structure and the customer service times. The static planning linear programming approach (Harrison and Lopez [22]) is used to obtain a maximum throughput compatibility tree and to show that FCFS using this compatibility tree is throughput optimal. We investigate matching rates and show by Hotelling's  $T^2$ -test and simulation that they are dependent on the service time distribution.

*Key words:* parallel service system, deterministic fluid approximation; matching rate.

## 1 Introduction

Parallel service systems are widely used to model service and manufacturing systems. Such systems have parallel servers  $\mathcal{S} = \{s_1, \dots, s_J\}$  of various skills, a stream of customers of various types  $\mathcal{C} = \{c_1, \dots, c_I\}$ , and a bipartite compatibility graph  $\mathcal{G}$  where  $(s_j, c_i) \in \mathcal{G}$  if server  $s_j$  can serve customers of type  $c_i$ ; see Figure 1. In this paper we focus on the behavior of such systems under the policy of first come first served (FCFS), and

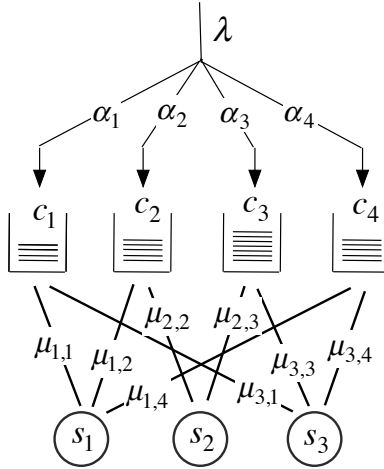


Figure 1: A parallel skilled based service system with 3 servers and 4 customer types

in particular, on deterministic fluid approximations for such systems uniformly scaled by time and space.

It is well known that the policy of FCFS for parallel service systems is not optimal, in that it may waste resources and result in longer waiting times than under other policies. It is nevertheless very widely used, because it is simple to implement, does not require any knowledge of system parameters, and is fair to

\*Department of Statistics, The University of Haifa, Mount Carmel 31905, Israel, [yuval@stat.haifa.ac.il](mailto:yuval@stat.haifa.ac.il). Research supported in part by Israel Science Foundation Grant 286/13.

†Department of Statistics, The University of Haifa, Mount Carmel 31905, Israel, [gweiss@stat.haifa.ac.il](mailto:gweiss@stat.haifa.ac.il). Research supported in part by Israel Science Foundation Grant 286/13.

‡Department of Decision Sciences, School of Business, National University of Singapore, Singapore, [bizzhq@nus.edu.sg](mailto:bizzhq@nus.edu.sg).

customers. An important property of FCFS is as follows: Assume that arriving customers have complete information of the system at their arrival, and can choose among the compatible servers which queue to join, and each server uses FCFS for his queue. In that case, the policy of join the shortest work load (JSW) will be the Nash equilibrium for customers that wish to minimize their waiting times. But this policy of JSW is automatically achieved when customers queue up in a single queue and the servers are using FCFS. FCFS can then serve as a benchmark, and comparison with other policies will provide an estimate of the “price of anarchy”. In particular, performance under FCFS may help in designing the system, e.g., deciding on an improved compatibility graph, and improved service rates.

Moreover, using FCFS has two purposes: under appropriate conditions it introduces resource pooling, i.e., all servers are busy at the same time and act like a combined server, and it gives the same service level to customers of all different types, i.e., it achieves approximately global FCFS (as defined by Talreja and Whitt [28]).

Our goal in this study is to determine conditions for complete resource pooling, i.e., conditions on the system parameters such that under FCFS all the servers can act as a combined server providing global FCFS, no matter what the arrival rate is, and to determine the maximal service capacity of the system in that case. This maximal service capacity determines stability under any arrival rates, including time-varying arrival rates. Alternatively, if complete resource pooling fails to hold, we wish to determine whether the servers decompose uniquely to subsets that have complete resource pooling.

The literature on parallel service system is quite voluminous. An incomplete list would include an early study [18]; applications to manufacturing and supply chain management [26, 31], applications to call centers and internet service systems [16, 21, 27, 33], attempts to find optimal policies, mainly for small graph systems [7, 8, 9, 17, 29, 34], heavy traffic and fluid approximations [22, 23], and many-server scaling [1, 19, 20]. In view of [20, 22, 23], establishing fluid approximations is often the first step to solve the optimal dynamic control problem for parallel service systems. *Thus, it would be necessary to provide a unified framework of establishing fluid approximations for such system with an arbitrary compatibility graph (topology).*

On the other hand, to evaluate the utilization of each server and customer quality of service, one needs to compute matching rates: the fraction of services by server  $s_j$  to customers of type  $c_i$ , out of all services performed by the system. It is straightforward to see that the matching rates immediately determine resource pooling and maximal service capacity. Adan and Weiss [5] discuss the special case when service rates depend only on the server, arrivals are Poisson, and service is exponential, under the policy of FCFS-ALIS (assign longest idle server) and derive a product-form stationary distribution for this system. From the stationary distribution it is possible to calculate matching rates, which in heavy traffic are equal to those obtained for the FCFS infinite bipartite matching model of [2, 4, 11]. The matching rates of the FCFS infinite bipartite matching model reappear in the analysis of parallel service systems with many servers, as demonstrated empirically in [1]. *Motivated by Adan and Weiss [5], we want to see whether the matching rates can be completely determined from the first moments of the customers interarrival and service times, in general, or under specific assumptions on the topology of the system and the interarrival and service distributions.*

Furthermore, as observed from some of the above literature, to understand the behavior of parallel service systems one needs to characterize the dynamics of the positions of the  $J$  servers in the queue. The position dynamics of the  $J$  servers can be used to determine whether resource pooling holds and to calculate the maximal service capacity of the system. Adan and Weiss [4, 5] characterize the position dynamics of the servers in the case of server dependent service rates, Poisson arrivals and exponential service times. *The natural question is whether we can determine the fluid trajectories of the positions of the  $J$  servers under general assumptions on customer arrivals and service times.*

Finally, by the work of Dai [12] for multiclass queueing networks, fluid approximations not only provide an asymptotic analysis but also verify stability in the sense of positive Harris recurrence and existence of stationary distributions. *One would like to see whether fluid approximations can also be used to verify the stability for parallel service systems.* Foss and Chernova [15] consider parallel service systems under JSW (as well as join shortest queue, JSQ). They derive conditions for stability when the service rates depend only on the servers and not on the customer types, and also when the service rates depend only on the customer type and not on the server. For the general case, when service rates depend both on the server and customer type, they produce an example in which stability depends not only on service rates but also on the complete distributions of the service times — this means that the fluid model is not informative enough to determine stability. *Thus it would be interesting to find conditions such that the system stability can be determined by*

the corresponding fluid approximations.

Mainly motivated by the above, in this paper we focus on the following questions

- Establish the deterministic fluid approximations;
- Explore stability conditions for parallel service systems using the fluid model approach;
- Obtain explicit fluid trajectories for the server-dependent (SD) and customer-dependent (CD) processing rates cases;
- Find matching rates for parallel service systems with complete bipartite graphs, tree bipartite graphs, or hybrids of those;
- Derive a bound on service capacity, and obtain an optimal tree graph for which FCFS achieves the bound and is throughput optimal, by introducing and solving an LP static planning problem;
- Demonstrate by Hotelling's  $T^2$ -test and simulation of the SD case that matching rates depend on the service time distributions.

The rest of the paper is organized as follows. In Section 2 we describe our model, define stochastic processes that describe its dynamics, and define matching rates and resource pooling. Section 3 is devoted to the fluid model. We show that fluid limits exist, and derive some fluid model equations that every fluid limit needs to satisfy. The discussion on the stability of parallel service systems and an example of Foss and Chernova [15] are given in Section 4. In Section 5, we use the fluid model equations to obtain the explicit fluid trajectories for the SD (server-dependent) processing rates case. The fluid model equations are also used to obtain the explicit fluid trajectories for the CD (customer-dependent) processing rates case in Section 6. We obtain fluid trajectories for parallel service systems with complete bipartite graphs, tree bipartite graphs, or hybrids of those, by calculating matching rates in Section 7. We formulate an LP static planning problem (cf. [22]) to find a bound on service capacity, and obtain an optimal tree graph for which FCFS achieves the bound and is throughput optimal in Section 8. Finally in Section 9 we demonstrate by simulation of the SD case, that matching rates depend on the service time distributions, but are very close to the values computed analytically for exponential service times.

## 2 The stochastic system model

Given the servers  $\{s_1, \dots, s_J\}$ , the customer types  $\{c_1, \dots, c_I\}$  and the compatibility graph  $\mathcal{G}$ , the primitives of the stochastic system consist of a sequence of interarrival times, a sequence of customer types, and one sequence of processing times for each compatibility link in the graph  $\mathcal{G}$ . We assume all these sequences are independent.

We let  $a(\ell)$  be the arrival time of the  $\ell$ th customer and  $u(\ell) = a(\ell) - a(\ell - 1)$  be the interarrival times, where  $\ell = 0, \pm 1, \pm 2, \dots$ , and  $a(0) \leq 0 < a(1)$ , and we let  $A(t) = \max\{\ell : a(\ell) \leq t\}$ . The distribution of  $u(\ell)$  is  $F$  with mean  $1/\lambda$ , so that  $A(t)$  is a renewal process with rate  $\lambda$  (all the fluid model results below continue to hold if we assume only that the arrival process  $A(t)$  is stationary and  $A(t)/t \rightarrow \lambda$  a.s.). In particular, for  $s < t$ ,  $A(t) - A(s)$  counts the total number of arrivals in  $(s, t]$ . Customer types are i.i.d., type  $c_i$  has probability  $\alpha_{c_i}$ ,  $i = 1, \dots, I$ , and we let  $\xi(\ell)$  be a unit vector of length  $I$  such that  $\xi_i(\ell) = 1$  if customer  $\ell$  is of type  $c_i$ , for  $\ell = 0, \pm 1, \pm 2, \dots$ . The counts of arrivals of customers of each type are then given by

$$A_{c_i}(t) = \begin{cases} \sum_{\ell=1}^{A(t)} \xi_i(\ell), & t \geq 0, \\ - \sum_{\ell=A(t)+1}^0 \xi_i(\ell), & t < 0 \end{cases} \quad (1)$$

We let  $v_{s_j, c_i}(0)$  be the remaining service time of server  $s_j$  if he is serving a customer of type  $c_i$  at time 0, and  $v_{s_j, c_i}(0) = 0$  otherwise. We let  $v_{s_j, c_i}(k)$ ,  $k = 1, 2, \dots$ , be the processing time of the  $k$ th customer of type  $c_i$  that server  $s_j$  is serving after time 0. The distribution of  $v_{s_j, c_i}(k)$ ,  $k = 1, 2, \dots$ , is  $G_{s_j, c_i}$  with mean

$m_{s_j, c_i}$  and rate  $\mu_{s_j, c_i} = 1/m_{s_j, c_i}$ . We let  $X_{s_j, c_i}(t) = \max\{k : \sum_{\ell=0}^k v_{s_j, c_i}(\ell) \leq t\}$  count the number of job completions by server  $s_j$  when processing customers of type  $c_i$  for a total processing time  $t$ , so that  $X_{s_j, c_i}(t)$  is a renewal process of rate  $\mu_{s_j, c_i}$  (all the fluid model results continue to hold if we assume only that the service completion process  $X_{s_j, c_i}(t)$  is stationary and  $X_{s_j, c_i}(t)/t \rightarrow \mu_{s_j, c_i}$  a.s.).

Service policy is FCFS, i.e., when a server becomes available he will next serve the compatible customer that has been waiting for the longest time. To complete the service policy description, when a customer arrives and there are several idle compatible servers, then the customer is assigned to the compatible server that has been idle for the longest time, i.e., assign longest idle server, ALIS.

The special case when service rates depend only on the servers, and when arrivals are Poisson and services are exponentially distributed is tractable, and is analyzed in [5] (see also [32]). Under these assumptions the system can be described by a countable state continuous time Markov chain, and most surprisingly, this Markov chain has a product form stationary distribution. The following Figure 2 describes the state of the Markov chain: The circles represent the customers in the system ordered from left to right by order of arrivals,  $\ell$  busy servers are placed with the customers which they are currently serving, followed by  $J - \ell$  idle servers ordered by their idleness times, so that  $M_1, \dots, M_J$  is a permutation of the servers  $s_1, \dots, s_J$ . The state at time  $t$  is defined as  $\mathcal{X}(t) = (M_1, n_1, \dots, M_\ell, n_\ell, M_{\ell+1}, \dots, M_J)$ , where  $n_j$  counts the number of customers queueing between servers  $M_j$  and  $M_{j+1}$ . All the customers between  $M_j$  and  $M_{j+1}$  have been skipped by servers  $M_{j+1}, \dots, M_J$  and must therefore be of types in the set  $\mathcal{U}(M_1, \dots, M_j)$  of customer types which are unique customers of  $M_1, \dots, M_j$ , where  $\mathcal{U}(M_1, \dots, M_j)$  is the set of customer types who are not compatible with servers  $\mathcal{S} \setminus \{M_1, \dots, M_j\}$  (see the definition at the end of this section). The dynamics are as follows:

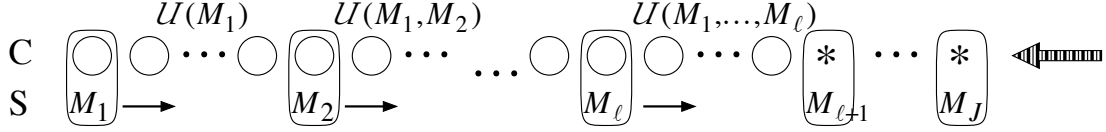


Figure 2: A state for the Markovian FCFS-ALIS parallel skill based system

Customers arrive from the right, scan the idle servers and join the end of the queue with the first compatible idle server that they find, or without a server if none is available. When a server completes a service, a customer leaves the system, and the server moves to the right, scanning the waiting customers until he finds the first compatible customer, or if no such customer is available, he joins the end of the idle servers queue at its left end. Under the assumption that service rates depend only on the server, Poisson arrivals and exponential services, this is a discrete state continuous time Markov chain.

$\mathcal{X}(t) = (M_1, n_1, \dots, M_i, n_i, M_{i+1}, \dots, M_J)$  in itself is not a Markov process for our general system, but if we add the remaining time to the next arrival and the remaining times until service completion for all busy servers, it becomes a Markov process in continuous time with an uncountable state space. In this paper we consider the dynamics of this more general system, and study its fluid limits. To describe the dynamics we use a more detailed representation of the system as illustrated in Figure 3, where the system has 3 servers, 3 customer types, and the compatibility graph includes  $\mathcal{G} = \{(s_1, c_1), (s_2, c_1), (s_2, c_2), (s_3, c_1), (s_3, c_3)\}$ . On the horizontal time axis the arrival times of customers are marked by  $a(\ell)$ . On the vertical axis the types of successive customers are listed. The list includes all the customers, past present and future, starting from the oldest customer that was present at time 0. For each customer there is a horizontal line starting at his arrival, and ending at his departure, which includes his waiting time and his service time. With each of the  $J$  servers there is a path that describes his whole history, composed of horizontal intervals when he is serving a customer, and vertical intervals that connect the end of service of a customer and the beginning of service of the next customer that he is serving. A top path describes the counting process of the arrival stream  $A(t)$ . When a server is idle he will move together with  $A(t)$ .

Our working hypothesis is that if we scale time and space uniformly by  $n$  and let  $n$  increase, we will get fluid limits which will evolve along piecewise linear paths, so that the fluid limits of the picture in Figure 3 will look as in Figure 4. Here the horizontal and vertical steps of the paths of servers become increasing straight lines. The top line records all cumulative fluid arrivals as a function of time, and the arriving fluid is a mixture of the three customer types. Under the line of server  $s_3$  all the arrivals have already departed. In the area between the lines of servers  $s_2, s_3$  fluid of customers of types  $c_3$  are still waiting, but types  $c_1, c_2$



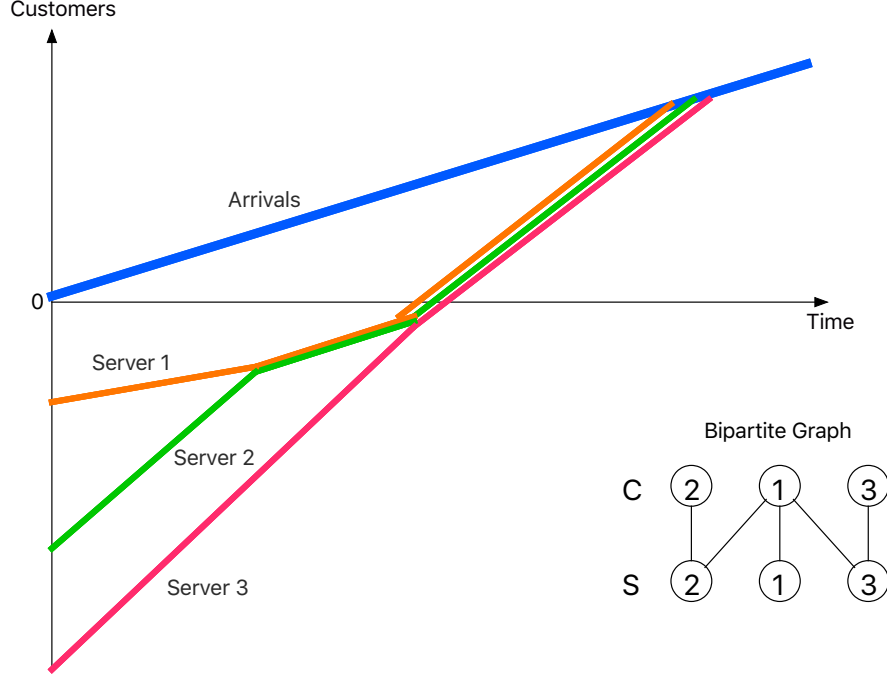


Figure 4: Conjectured Fluid Dynamics under FCFS-ALIS

have already departed. In the area between the lines of server  $s_1$  and  $s_2$  only customer fluid of type  $c_1$  are departed, and customers of types  $c_2, c_3$  are still waiting. Finally, between the arrival line and the line of server  $s_1$  fluid customers of all three types are still waiting. In this figure all three lines eventually meet; this is the phenomena of *complete resource pooling*. Furthermore, in this instance the fluid limit is stable, as all fluid is drained and the fluid system is empty from some time onwards.

Assuming that fluid limits move along such straight lines, we are interested in the following questions:

- When  $\lambda$  is large, do the lines of all the servers merge eventually? If so, we say that complete resource pooling holds.
- If the lines do not merge, does this define a unique decomposition of the servers?
- For what values of  $\lambda$  do all the lines eventually merge with  $\lambda t$ , the top line? In cases when they merge, we say that the fluid model is stable.

Complete resource pooling implies under some minor conditions that queues between servers are stable, and stability of the fluid model implies under some minor conditions that the stochastic system is stable.

We introduce some notation: We denote by  $\mathcal{C}(s_j)$  the customer types compatible with  $s_j$ , referred to as customers of  $s_j$ , and by  $\mathcal{S}(c_i)$  the servers that are compatible with customers of type  $c_i$ , referred to as the servers of  $c_i$ . For a subset  $C \subseteq \mathcal{C}$  of customer types we let  $\mathcal{S}(C) = \bigcup_{c_i \in C} \mathcal{S}(c_i)$  denote all the servers of customer types in  $C$ . Also, for a subset  $S \subseteq \mathcal{S}$  we let  $\mathcal{C}(S) = \bigcup_{s_j \in S} \mathcal{C}(s_j)$  denote all the customer types that can be served by some servers in  $S$ , and we let  $\mathcal{U}(S) = \overline{\mathcal{C}(S)}$  denote the set of customer types which cannot be served by any server outside  $S$ , that is, the unique customers of  $S$ . For a subset  $C \subseteq \mathcal{C}$  of customer types we let  $\alpha_C = \sum_{c_i \in C} \alpha_{c_i}$ .

To describe the dynamics of the system we define the following quantities:

$P_{s_j}(t)$  is the position of server  $s_j$  at time  $t$ , where we let  $P_{s_j}(t) = \ell$  if the server is serving at time  $t$  the  $\ell$ th customer in the sequence of arrivals. If servers  $s_{j_1}, \dots, s_{j_k}$  are idle at time  $t$  then their positions are defined as  $A(t) + 1, \dots, A(t) + k$ , ordered by duration of idleness, with  $A(t) + k$  the longest idle.

$Y_j(t)$  is the current  $j$ th level, where we let  $Y_1(t) < \dots < Y_J(t)$  be the ordered set of the positions of the servers at time  $t$ .

$M_1(t), \dots, M_J(t)$  is the random permutation of the servers at time  $t$ , where we let  $P_{M_j(t)}(t) = Y_j(t)$

$T_{s_j, c_i}(t)$  is the cumulative time over  $(0, t)$  that server  $s_j$  has served customers of type  $c_i$ .

In this paper we will mainly investigate the processes  $Y_j(t), M_j(t)$ ,  $j = 1, \dots, J$ . These processes also define the actual queue lengths. We let  $Q_{c_i, j}(t)$  denote the number of customers of type  $c_i$  which are waiting between servers  $M_j(t)$  and  $M_{j+1}(t)$  at time  $t$ . These are given by:

$$Q_{c_i, j}(t) = \begin{cases} \sum_{\ell=Y_j(t)+1}^{Y_{j+1}(t)-1} \xi_i(\ell) \mathbf{1}\{c_i \in \mathcal{U}(M_1(t), \dots, M_j(t))\}, & j = 1, \dots, J-1, \\ \sum_{\ell=Y_J(t)+1}^{A(t)} \xi_i(\ell), & j = J. \end{cases} \quad (2)$$

where  $\mathbf{1}\{\cdot\}$  is the indicator function.

Let  $U(t)$  be the remaining time at time  $t$  until next arrival,  $V_{s_j, c_i}(t)$  be the remaining processing time of  $c_i$  by  $s_j$  if  $s_j$  is processing a type  $c_i$  customer at time  $t$ , and  $V_{s_j, c_i}(t) = 0$  otherwise. The initial state of the system is given by  $A(0) = 0$ ,  $P_{s_j}(0)$ ,  $U(0) = a(1)$ , and  $V_{s_j, c_i}(0) = v_{s_j, c_i}(0)$  (note that  $P_{s_j}(0) < 0$ ). Both  $\mathcal{Y}(t) = (A(t), P_{s_j}(t), U(t), V_{s_j, c_i}(t))$ , and  $\mathcal{X}(t) = (M_j(t), Q_{c_i, j}(t), U(t), V_{s_j, c_i}(t))$  are Markov processes. The former is always transient, as  $A(t), P_{s_j}(t)$  are non-decreasing with  $t$ . The latter may be stable, and we say that the queueing system is stable (ergodic) if  $\mathcal{X}(t)$  is positive Harris recurrent (ergodic).

### 3 Fluid limits and fluid model equations

To study the fluid limits of the system we consider a sequence of systems defined over the same probability space, indexed by  $n = 1, 2, \dots$ , and study their fluid scaling. All the systems in the sequence share the same stochastic sequences  $A(t), \xi(\ell), X_{s_j, c_i}(t)$ , but they differ in their initial conditions: We let  $P_{s_j}^n(0)$ ,  $j = 1, \dots, J$  be the initial positions of the servers in the  $n$ th system. We denote quantities of the  $n$ th system which are not common to all systems by the superscript  $n$ . We obtain the fluid scaling for the sequence of systems by scaling time and space of the  $n$ th system by  $n$ . For any function  $z^n(t)$  we define the fluid scaling as  $\bar{z}^n(t) = \frac{1}{n} z^n(nt)$ .

Consider sample paths  $\omega \in \Omega$  of the sequence of systems, and consider one of the processes, say  $z^n(t, \omega)$ . If  $\bar{z}^r(t, \omega) = \frac{1}{r} z^r(rt, \omega) \rightarrow \bar{z}(t)$  uniformly on compacts (u.o.c.) when  $r \rightarrow \infty$ , for some  $\omega$  and for some subsequence  $r$  of  $n = 1, 2, \dots$ , where  $\bar{z}(t)$  is a deterministic function of  $t$ , then we say that  $\bar{z}(t)$  is a fluid limit of  $z^n(\cdot, \cdot)$ .

To obtain the fluid dynamics of our system we need to assume that the following holds:

$$\lim_{n \rightarrow \infty} \bar{P}_{s_j}^n(0) = \lim_{n \rightarrow \infty} \frac{P_{s_j}^n(0)}{n} = \bar{P}_{s_j}(0) \leq 0. \quad (3)$$

$$\lim_{n \rightarrow \infty} \frac{U^n(0)}{n} = 0, \quad \lim_{n \rightarrow \infty} \frac{V_{s_j, c_i}^n(0)}{n} = 0. \quad (4)$$

and

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} A(nt, \omega) &= \lambda t \quad \text{u.o.c., a.s.} \\ \lim_{n \rightarrow \infty} \frac{1}{n} A_{c_i}(nt, \omega) &= \lambda \alpha_{c_i} t \quad \text{u.o.c., a.s.} \\ \lim_{n \rightarrow \infty} \frac{1}{n} X_{s_j, c_i}(nt, \omega) &= \mu_{s_j, c_i} t \quad \text{u.o.c., a.s.} \end{aligned} \quad (5)$$

We assume throughout that (3) and (4) hold. Assumptions (5) hold for our system by the functional strong law of large numbers, since we assume renewal arrivals, i.i.d. customer types, and renewal service times. We exclude the set of measure zero where (5) fails to hold. Let  $T_{s_j, c_i}^n(t)$  be the cumulative service time of customer type  $c_i$  provided by server  $s_j$  over time interval  $[0, t]$ . The following theorem proves the existence of fluid limits.

**Theorem 1.** *Fluid limits for  $\bar{T}_{s_j, c_i}^n(t, \omega)$ ,  $\bar{P}_{s_j}^n(t, \omega)$ ,  $\bar{Y}_j^n(t, \omega)$ ,  $\bar{Q}_{c_i, j}^n(t)$  exist almost surely for every  $\omega$ , and they are almost surely Lipschitz continuous for every  $t > 0$ .*

*Proof.* Consider first  $T_{s_j, c_i}^n(t, \omega)$ . We have for all  $\omega$  that  $T_{s_j, c_i}^n(t, \omega) - T_{s_j, c_i}^n(s, \omega) \leq t - s$  for all  $s < t$ , and so for all  $\omega$  and every  $n$  also  $\bar{T}_{s_j, c_i}^n(t, \omega) - \bar{T}_{s_j, c_i}^n(s, \omega) \leq t - s$ , so the sequence is equicontinuous and uniformly bounded on every compact interval, for every  $\omega$ . Fix  $\omega$ . By Arzela-Ascoli theorem, for every compact interval there exists a subsequence  $r$  of  $n$  such that  $\bar{T}_{s_j, c_i}^r(t, \omega)$  converges to some  $\bar{T}_{s_j, c_i}(t)$  as  $r \rightarrow \infty$  uniformly on the interval. It is then possible to choose a further subsequence that will converge uniformly on all compacts. Furthermore, all  $\bar{T}_{s_j, c_i}^n(t, \omega)$  are Lipschitz continuous for every  $\omega$ , and hence so is every fluid limit  $\bar{T}_{s_j, c_i}(t)$ .

The main part of the proof is to show the existence of fluid limits for  $P_{s_j}^n(t, \omega)$ . The functions  $P_{s_j}^n(t, \omega)$  are non-decreasing in  $t$ , and  $P_{s_j}^n(0, \omega) \leq P_{s_j}^n(t, \omega) \leq A(t, \omega) + J$ , and hence for any  $\epsilon > 0$  and large enough  $n$ ,  $\bar{P}_{s_j}(0) - \epsilon \leq \bar{P}_{s_j}^n(t, \omega) \leq \lambda t + \epsilon$ , so  $\bar{P}_{s_j}^n(t, \omega)$  are non-decreasing and uniformly bounded at each  $t$  for all  $n$ . Hence we can find a subsequence  $r$  such that  $\bar{P}_{s_j}^r(t, \omega) \rightarrow \bar{P}_{s_j}(t)$  as  $r \rightarrow \infty$  for all rational  $t$ , and we have that  $\bar{P}_{s_j}(t)$  is non-decreasing on all rationals, and we can then extend its definition to all real  $t$ . If we can show that  $\bar{P}_{s_j}(t)$  is continuous, then by Lemma 4.1 of Dai [12] we will have that  $\bar{P}_{s_j}^r(t, \omega) \rightarrow \bar{P}_{s_j}(t)$  uniformly on compacts. We will show that  $\bar{P}_{s_j}(t)$  is in fact Lipschitz continuous for  $t > 0$ .

We note that  $\bar{P}_{s_j}(t)$  may be discontinuous at 0. Consider the limiting  $\bar{P}(0), \bar{Y}(0), \bar{M}(0)$ , and assume the following: (i)  $\bar{Y}_k(0) = \bar{P}_{s_j}(0)$  (ii)  $\mathcal{C}(s_j) \subseteq \mathcal{C}(\bar{M}_{k+1}(0), \dots, \bar{M}_J(0))$  (iii)  $\bar{Y}_k(0) < \bar{Y}_{k+1}(0)$ . Denote  $v_{s_j}(0, \omega) = \max_{c_i \in \mathcal{C}(s_j)} v_{s_j, c_i}(0, \omega) > 0$ . Then we have that  $P_{s_j}^n(v_{s_j}(0, \omega)) > Y_{k+1}^n(0)$ , and so we have:

$$\lim_{n \rightarrow \infty} \bar{P}_{s_j}^n(0, \omega) = \bar{Y}_k(0) \quad \text{while} \quad \bar{P}_{s_j}(0+) \geq \liminf_{n \rightarrow \infty} \frac{1}{n} P_{s_j}^n\left(\frac{1}{n} v_{s_j}(0, \omega)\right) \geq \bar{Y}_{k+1}(0) > \bar{Y}_k(0).$$

Consider now  $v_{s_j}^n(0, \omega) < t_0 < t_1$ . Let  $c_i$  be the type of the customer that  $s_j$  is serving at time  $t_0$ , let  $w_{s_j}(t_0, \omega)$  be elapsed time of this customer, and let  $t = t_0 - w_{s_j}(t_0, \omega)$  be the time at which the service of this customer started. At time  $t$ , by FCFS, all customers of type  $c_i$  in positions  $> P_{s_j}^n(t, \omega)$  have not yet started service. During the time period  $(t, t_1)$ , server  $s_j$  is processing customers of type  $c_i$  as well as customers of types  $c_l \in \mathcal{C}(s_j)$ ,  $c_l \neq c_i$ . Hence it may only process at most  $X_{s_j, c_i}(T_{s_j, c_i}^n(t_1)) - X_{s_j, c_i}(T_{s_j, c_i}^n(t))$  customers of type  $c_i$ . During the time period  $(t, t_1)$ , other servers  $s_k \in \mathcal{S}(c_i)$ ,  $s_k \neq s_j$  may also be processing customers of type  $c_i$ . Therefore the total number of customers of types  $c_i$  that may be processed in the time period  $(t, t_1)$  cannot exceed  $\sum_{k \in \mathcal{S}(c_i)} (X_{s_k, c_i}(T_{s_k, c_i}^n(t_1)) - X_{s_k, c_i}(T_{s_k, c_i}^n(t)))$ .

We repeat the argument of the last paragraph for the scaled processes. Consider any  $0 < t_0 < t_1$  and  $n$  large enough that  $v_{s_j}^n(0) < nt_0$  almost surely by (4). Assume server  $s_j$  is working on job type  $c_i$  at time  $nt_0$ , with elapsed time  $w_{s_j}^n(nt_0, \omega)$ , and let  $nt = nt_0 - w_{s_j}^n(nt_0, \omega)$  be the time that processing of this job started. Then:

$$\begin{aligned} & \frac{1}{n} \left( P_{s_j}^n(nt_1, \omega) - P_{s_j}^n(nt_0, \omega) \right) \\ &= \frac{1}{n} \left( P_{s_j}^n(nt_1, \omega) - P_{s_j}^n(nt, \omega) \right) \\ &= \frac{\sum_{k=1}^I \sum_{\ell=P_{s_j}^n(nt, \omega)}^{P_{s_j}^n(nt_1, \omega)} \xi_k(\ell)}{\sum_{\ell=P_{s_j}^n(nt, \omega)}^{P_{s_j}^n(nt_1, \omega)} \xi_i(\ell)} \frac{1}{n} \sum_{\ell=P_{s_j}^n(nt, \omega)}^{P_{s_j}^n(nt_1, \omega)} \xi_i(\ell) \\ &\leq \frac{\sum_{k=1}^I \sum_{\ell=P_{s_j}^n(nt, \omega)}^{P_{s_j}^n(nt_1, \omega)} \xi_k(\ell)}{\sum_{\ell=P_{s_j}^n(nt, \omega)}^{P_{s_j}^n(nt_1, \omega)} \xi_i(\ell)} \frac{1}{n} \sum_{s_k \in \mathcal{S}(c_i)} \left( X_{s_k, c_i}(T_{s_k, c_i}^n(nt_1)) - X_{s_k, c_i}(T_{s_k, c_i}^n(nt)) \right) \end{aligned}$$

Going to the limit, we take a subsequence  $r$  for which convergence of  $\frac{1}{r} P_{s_j}^r(rt, \omega)$  holds. In the case that  $\lim_{r \rightarrow \infty} (P_{s_j}^r(rt_1, \omega) - P_{s_j}^r(rt_0, \omega)) < \infty$ , we have  $\bar{P}_{s_j}(t_1) - \bar{P}_{s_j}(t_0) = 0$ . Otherwise we now consider the above



inequality for all  $c_i$ . We have that as  $r \rightarrow \infty$ :

$$\begin{aligned}
\bar{P}_{s_j}(t_1) - \bar{P}_{s_j}(t_0) &\leq \max_{c_i \in \mathcal{C}} \left[ \lim_{r \rightarrow \infty} \left\{ \frac{\sum_{k=1}^I \sum_{\ell=P_{s_j}^r(rt_1, \omega)}^{P_{s_j}^r(rt_1, \omega)} \xi_k(\ell)}{\sum_{\ell=P_{s_j}^r(rt_1, \omega)}^{P_{s_j}^r(rt_1, \omega)} \xi_i(\ell)} \right\} \right. \\
&\quad \left. \times \lim_{r \rightarrow \infty} \left\{ \frac{1}{r} \sum_{s_k \in \mathcal{S}(c_i)} \left( X_{s_k, c_i}(T_{s_k, c_i}^r(rt_1)) - X_{s_k, c_i}(T_{s_k, c_i}^r(rt_0)) \right) \right\} \right] \\
&= \max_{c_i \in \mathcal{C}} \left[ \frac{1}{\alpha_{c_i}} \sum_{s_k \in \mathcal{S}(c_i)} \mu_{s_k, c_i} \left( \bar{T}_{s_k, c_i}(t_1) - \bar{T}_{s_k, c_i}(t_0) \right) \right] \\
&\leq \max_{c_i \in \mathcal{C}} \left[ \frac{1}{\alpha_{c_i}} \sum_{s_k \in \mathcal{S}(c_i)} \mu_{s_k, c_i} \right] \left( t_1 - t_0 + \lim_{r \rightarrow \infty} \frac{w_{s_j}^r(rt_0, \omega)}{r} \right) \\
&= \max_{c_i \in \mathcal{C}} \left[ \frac{1}{\alpha_{c_i}} \sum_{s_k \in \mathcal{S}(c_i)} \mu_{s_k, c_i} \right] (t_1 - t_0).
\end{aligned}$$

The last equality holds because  $\max_{1 \leq \ell \leq n} V_{s_j, c_i}^n(\ell)/n \rightarrow 0$  as  $n \rightarrow \infty$  for all  $s_j, c_i$  a.s. for all  $\omega$ . We have therefore that the fluid limits  $\bar{P}_{s_j}(t)$  are Lipschitz continuous with constant  $\max_{c_i \in \mathcal{C}} \left[ \frac{1}{\alpha_{c_i}} \sum_{k \in \mathcal{S}(c_i)} \mu_{s_k, c_i} \right]$ .

We can now use subsequences of subsequences to obtain that  $\bar{P}_{s_j}^r(t, \omega) \rightarrow \bar{P}_{s_j}(t)$  u.o.c. for all  $s_j$  as  $r \rightarrow \infty$ . For this subsequence we then have that  $\bar{Y}_j^r(t, \omega) \rightarrow \bar{Y}_j(t)$  which are the ordered values of  $\bar{P}_{s_j}(t)$ .

Finally, from (2) we obtain that for almost all  $\omega$  there is some subsequence  $r$  such that as  $r \rightarrow \infty$ :

$$\bar{Q}_{c_i, j}^r(t, \omega) \rightarrow \begin{cases} \alpha_{c_i}(\bar{Y}_{j+1}(t) - \bar{Y}_j(t)), & c_i \in \mathcal{U}(M_1(t), \dots, M_J(t)), j = 1, \dots, J-1, \\ \alpha_{c_i}(\lambda t - \bar{Y}_J(t)), & c_i \in \mathcal{C}, j = J. \end{cases} \quad (6)$$

□

We now know that almost surely for all  $\omega$  there exist subsequences which lead to fluid limits that are Lipschitz continuous for all  $t > 0$ . We also assume that for these subsequences (3)-(5) hold by excluding a set of measure zero. Since the fluid limits are Lipschitz continuous they are absolutely continuous and hence possess derivatives almost everywhere, and are integrals of their derivatives. We shall call times  $t$  at which derivatives of fluid limits exist regular times. We will use  $\dot{z}(t)$  to denote  $\frac{d}{dt} z(t)$  for all fluid limits, for all regular  $t$ . We now wish to derive equations which all fluid limits must satisfy almost surely.

By definition, for every  $n$ , at every time  $t$ ,  $\bar{P}_{s_j}^n(t, \omega)$  for  $s_1, \dots, s_J$  are all different, so that

$$\bar{Y}_1^n(t, \omega) = \frac{1}{n} P_{M_1^n(nt, \omega)}^n(nt, \omega) < \dots < \bar{Y}_J^n(t, \omega) = \frac{1}{n} P_{M_J^n(nt, \omega)}^n(nt, \omega)$$

However, for the fluid limits we only have that  $\bar{Y}_1(t) \leq \bar{Y}_2(t) \leq \dots \leq \bar{Y}_J(t)$ . As a result the fluid limits no longer define a unique permutation of the servers at time  $t$ , and instead we have an ordered partition of  $s_1, \dots, s_J$ . For concreteness we order  $\bar{P}_{\bar{M}_1(t)}(t), \dots, \bar{P}_{\bar{M}_J(t)}(t)$  so that  $\bar{P}_{\bar{M}_j(t)}(t) < \bar{P}_{\bar{M}_{j+1}(t)}(t)$  or  $\bar{P}_{\bar{M}_j(t)}(t) = \bar{P}_{\bar{M}_{j+1}(t)}(t)$  and  $\bar{M}_j(t) < \bar{M}_{j+1}(t)$ . We define the fluid ordered partition  $\bar{\mathcal{S}}(t) = (\bar{S}_1(t), \dots, \bar{S}_L(t))$  as follows:

$$\begin{aligned}
&(\bar{S}_1(t), \dots, \bar{S}_L(t)) \text{ is a partition of } \mathcal{S}, \\
&M_j, M_{j'} \in \bar{S}_\ell(t) \Rightarrow \bar{P}_{M_j}(t) = \bar{P}_{M_{j'}}(t), \\
&M_j \in \bar{S}_\ell(t) \text{ and } M_{j'} \in \bar{S}_{\ell+1}(t) \Rightarrow \bar{P}_{M_j}(t) < \bar{P}_{M_{j'}}(t).
\end{aligned} \quad (7)$$

Note that  $\bar{\mathcal{S}}(t), \bar{M}(t)$  are limits at time  $nt$  when  $n \rightarrow \infty$ , but they are not scaled in space, since they are discrete and finite. We introduce the notation  $\bar{Y}_S(t), \dot{\bar{Y}}_S(t)$  to denote the common value of  $\bar{P}_{M_j}(t), \dot{\bar{P}}_{M_j}(t)$ ,  $M_j \in S$ .

We now have the following theorem on the dynamics of the fluid model. We use the convention that  $\mu_{s_j, c_i} = \dot{T}_{s_j, c_i} = 0$  for  $(s_j, c_i) \notin \mathcal{G}$ .

**Theorem 2.** Consider a fluid limit in which servers at levels  $k, \dots, l$  move together for a while, i.e.,  $\bar{Y}_{k-1}(\tau) < \bar{Y}_k(\tau) = \dots = \bar{Y}_l(\tau) < \bar{Y}_{l+1}(\tau)$  (or if  $l = J$ ,  $\bar{Y}_l(\tau) < \lambda\tau$ ), for some  $k \leq l$  and for all  $s < \tau < t$ . Let  $\bar{S}(\tau) = (S'(\tau), \{M_k, \dots, M_l\}, S''(\tau))$  for the same range of  $\tau$ , where  $S'(\tau)$  and  $S''(\tau)$  are the subsets of servers preceding and succeeding  $M_k, \dots, M_l$  (the sets  $S'(\tau)$  and  $S''(\tau)$  may themselves consist of a further partition, but this is irrelevant here). Then a.s. all fluid limits at  $s < \tau < t$  must satisfy the following equations:

$$\sum_{c_i \in \mathcal{C}(M_j) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)} \dot{T}_{M_j, c_i}(\tau) = 1 \quad j = k, \dots, l, \quad (8)$$

$$\dot{\bar{Y}}_k(\tau) = \dots = \dot{\bar{Y}}_l(\tau) = \frac{1}{\alpha_{c_i}} \sum_{j=k}^l \mu_{M_j, c_i} \dot{T}_{M_j, c_i}(\tau), \quad c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J). \quad (9)$$

*Proof.* Consider a fluid limit of all the processes obtained for some  $\omega$  and subsequence  $r$ , for which the assumptions of the theorem hold.

By the continuity of  $\bar{P}_{s_j}(\tau)$  the sets  $S'(\tau), \{\bar{M}_k(\tau), \dots, \bar{M}_l(\tau)\}, S''$  are constant for all  $s < \tau < t$ , and  $\bar{S}(\tau)$  is well defined. If  $\bar{Y}_j(\tau) < \bar{Y}_{l+1}(\tau) \leq \lambda\tau$ ,  $s < \tau < t$ ,  $j = k, \dots, l$  then for  $r$  large enough  $Y_j^r(r\tau, \omega) < Y_{l+1}^r(r\tau, \omega) < A(r\tau, \omega)$ ,  $\tau \in (s, t)$ ,  $j = k, \dots, l$ , which implies that  $M_k^r, \dots, M_l^r = M_k, \dots, M_l$  are the same for all  $r$  large enough, that all the servers  $M_k, \dots, M_l$  are busy all the time between  $(rs, rt)$ , and the types of customers which  $M_j$  will be serving will be  $c_i \in \mathcal{C}(M_j) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)$ . It follows that

$$\sum_{c_i \in \mathcal{C}(M_j) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)} \frac{1}{r} (T_{M_j, c_i}^r(rt, \omega) - T_{M_j, c_i}^r(rs, \omega)) = t - s$$

and (8) follows. For the same  $s, t$  and large enough  $r$ , we have:

$$Y_k^r(rs, \omega) = \min_{k \leq j \leq l} P_{M_j}^r(rs, \omega), \quad Y_l^r(rs, \omega) = \max_{k \leq j \leq l} P_{M_j}^r(rs, \omega),$$

$$Y_k^r(rt, \omega) = \min_{k \leq j \leq l} P_{M_j}^r(rt, \omega), \quad Y_l^r(rt, \omega) = \max_{k \leq j \leq l} P_{M_j}^r(rt, \omega).$$

Consider for  $c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)$  the two counts:

$$N_1^r(\omega) = \sum_{\ell=Y_k^r(rs, \omega)+1}^{Y_k^r(rt, \omega)-1} \xi_i(\ell, \omega), \quad N_2^r(\omega) = \sum_{\ell=Y_k^r(rs, \omega)}^{Y_l^r(rt, \omega)} \xi_i(\ell, \omega)$$

These count customers of type  $c_i$  which are associated with the time interval  $(rs, rt)$ : every customer of type  $c_i$  which appears in the first count has started service and finished service within the time period  $(rs, rt)$ . The second count includes all the customers of type  $c_i$  which have departed in the time interval  $(rs, rt)$ , including some that started processing at an earlier time, and also those which have started service and not departed yet.

Compare this to

$$N_3^r(\omega) = \sum_{j=k}^l \left( X_{M_j, c_i}(T_{M_j, c_i}^r(rt, \omega), \omega) - X_{M_j, c_i}(T_{M_j, c_i}^r(rs, \omega), \omega) \right),$$

which counts all the service completions of jobs of type  $c_i$ , served by one of the servers  $M_k, \dots, M_l$ , during the time interval  $(rs, rt)$  (recall that  $T_{M_j, c_i}^r(rt) - T_{M_j, c_i}^r(rs)$  is the total time that server  $M_j$  is processing customers of type  $c_i$  within the time interval  $(rs, rt)$ ). We have that  $N_2^r(\omega) \geq N_3^r(\omega) \geq N_1^r(\omega)$ .

However,

$$\lim_{r \rightarrow \infty} \frac{1}{r} N_1^r(\omega) = \lim_{r \rightarrow \infty} \frac{1}{r} N_2^r(\omega) = \alpha_{c_i}(\bar{Y}_k(t) - \bar{Y}_k(s)) = \dots = \alpha_{c_i}(\bar{Y}_l(t) - \bar{Y}_l(s)),$$

while

$$\lim_{r \rightarrow \infty} \frac{1}{r} N_3^r(\omega) = \sum_{j=k}^l \mu_{M_j, c_i} (\bar{T}_{M_j, c_i}(t) - \bar{T}_{M_j, c_i}(s)),$$

and (9) follows.  $\square$

**Corollary 1.** *If  $\bar{Y}_{j-1}(t) < \bar{Y}_j(t) < \bar{Y}_{j+1}(t)$  then*

$$\dot{\bar{Y}}_j(t) = \left( \sum_{c_i \in \mathcal{C}(M_j) \setminus \mathcal{C}(M_{j+1}, \dots, M_J)} \alpha_{c_i} m_{M_j, c_i} \right)^{-1} \quad (10)$$

*Proof.* From (9) we have that  $m_{M_j, c_i} \alpha_{c_i} \dot{\bar{Y}}_j(t) = \dot{T}_{M_j, c_i}(t)$ , and summing over  $c_i \in \mathcal{C}(M_j) \setminus \mathcal{C}(M_{j+1}, \dots, M_J)$  and using (8) we obtain (10).  $\square$

We refer to equations (8)–(10) as fluid model equations.

## 4 Stability

We are interested in verifying the following properties of fluid limits:

**Definition 1.** *Denote  $|\bar{P}(0)| = -\sum_{j=1}^J \bar{P}_{s_j}(0)$ .*

- (i) *We say that the fluid model is stable if starting from any fixed  $|\bar{P}(0)| = 1$ , there exists  $t_0$  such that for almost surely every fluid limit  $\lambda t - \bar{Y}_1(t) = 0$  for all  $t > t_0$ .*
- (ii) *We say that the fluid model has complete resource pooling if for all values of  $\lambda$ , starting from any fixed  $|\bar{P}(0)|$ , there exists  $t_0$  such that for almost surely every fluid limit  $\bar{Y}_J(t) - \bar{Y}_1(t) = 0$  for all  $t > t_0$ .*
- (iii) *We say that the fluid model has complete weak resource pooling if for all values of  $\lambda$ , starting from any fixed  $|\bar{P}(0)|$ , there exists  $t_0$  such that for almost surely every fluid limit  $\bar{Y}_1(t) = \dots = \bar{Y}_J(t)$*

Complete weak resource pooling is the situation in which in the limit, all servers move eventually at the same rate, but not together. Graphically, this means that the straight lines denoting their limiting paths become parallel, but may never merge.

**Definition 2.** *We say that the fluid model of a system under some given policy is maximum throughput with processing rate  $\mu^*$  if the fluid model for the given policy is stable for all  $\lambda < \mu^*$ , and if the fluid model of the system is unstable for all  $\lambda > \mu^*$  under every policy.*

Complete resource pooling and stability of the fluid limits and fluid model have far-reaching consequences for the stochastic system if some technical conditions are satisfied. In particular, in the following three theorems we will make the technical assumption that in the state space of the Markov processes considered, every bounded set of states is uniformly small. For definition of uniformly small sets of states in a Markov process, see Bramson [10] or Meyn and Tweedie [25].

**Theorem 3.** *Consider the Markov process  $\mathcal{Z}(t)$  and define  $\sum_{j=1}^J \sum_{i=1}^I (Q_{c_i, j}(t) + V_{s_j, c_i}(t)) + U(t)$  as its norm. Assume that every bounded set of states is uniformly small. If the fluid model of the system is stable then the process  $\mathcal{Z}(t)$  is ergodic, i.e. it possesses a stationary distribution, and the distribution of its state at time  $t$  converges to this stationary distribution as  $t \rightarrow \infty$ .*

*Proof.* This follows immediately from the fundamental theorem of Dai [12] and its extension in the monograph of Bramson [10].  $\square$

**Theorem 4.** *For the Markov process  $\mathcal{Z}(t)$  define  $\sum_{j=1}^J \sum_{i=1}^I (Q_{c_i, j}(t) + V_{s_j, c_i}(t)) + U(t)$  as norm, and assume that every bounded set of states is uniformly small as in Theorem 3. Consider the process  $\mathcal{Z}^0(t)$  obtained from  $\mathcal{Z}(t)$  by the exclusion of the components  $U(t), Q_{c_i, j}(t)$ . If complete resource pooling of the fluid model holds, and if  $\bar{Y}_J(t) < \lambda t$ , then there exists a measure  $\nu^0$  on the state space of  $\mathcal{Z}^0(\cdot)$  such that as  $t \rightarrow \infty$  the distribution of  $\mathcal{Z}^0(t)$  converges to  $\nu^0$ .*

*Proof.* Consider the same system with infinite supply of work, i.e., there is always a queue of customers waiting behind the most advanced server, of types  $c_i \in \mathcal{C}$  i.i.d. with probabilities  $\alpha_{c_i}$ . Then in this new system  $\mathcal{Z}^0(t)$  is a Markov process, and with the norm  $\sum_{i=1}^I \left( \sum_{j=1}^{J-1} Q_{c_i, j}(t) + \sum_{j=1}^J V_{s_j(t), c_i}(t) \right)$  every bounded set of states is uniformly small. If complete resource pooling holds then the fluid model of the process  $\mathcal{Z}^0(t)$  for the unlimited supply of work system is stable. Hence, by the fundamental theorem of Dai [12], the process

is ergodic, with some stationary measure  $\nu^0$ . Returning to the original system, and the process  $\mathcal{Z}(t)$ , we have  $\mathcal{Z}(t) = (\mathcal{Z}^0(t), U(t), Q_{c_i, J}(t))$  where the process  $\mathcal{Z}(t)$  is transient because by  $\bar{Y}_J(t) < \lambda t$  we have  $Q_{c_i, J}(t) \rightarrow \infty$  as  $t \rightarrow \infty$  almost surely. However, the process  $\mathcal{Z}(t)$  exactly satisfies the conditions of the Lemma of Adan, Foss, Shneer and Weiss [3]. It follows that the distribution of  $\mathcal{Z}^0(t)$  converges to  $\nu^0$ .  $\square$

We discuss the meaning of this theorem. It says that under complete resource pooling if the arrival rate is high the queue in front of all the servers will grow linearly but the servers will stay close together and move at some joint average rate, so that the permutation of the servers and the queues of customers between them will tend to a stationary distribution.

We consider now the case that there is no complete resource pooling. Consider a partition  $(S_1, \dots, S_L)$ , let  $S_\ell = \{M_k, \dots, M_l\}$  be the servers in positions  $k, \dots, l$ . We denote by  $Q_\ell = (Q_{c_i, k}, \dots, Q_{c_i, l-1}, c_i = 1, \dots, I)$  the queues of customers between the servers of  $S_\ell$ , and by  $v_\ell = (v_{M_k, c_i}, \dots, v_{M_l, c_i}, i = 1, \dots, I)$  the remaining processing times of the servers of  $S_\ell$ .

**Theorem 5.** Assume that for all  $t > 0$  there is a fixed partition  $(S_1, \dots, S_L)$  such that  $\bar{Y}_{S_1}(t) < \dots < \bar{Y}_{S_L}(t) < \lambda t$  and  $\dot{Y}_{S_1}(t) < \dots < \dot{Y}_{S_L}(t) < \lambda$ . Consider the processes  $\mathcal{Z}^\ell(t) = (M_j(t), Q_\ell(t), v_\ell(t) : M_j \in S_\ell)$ . Then there exist measures  $\nu^\ell$  on the state spaces of  $\mathcal{Z}^\ell(\cdot)$  such that as  $t \rightarrow \infty$  the distribution of  $\mathcal{Z}^\ell(t)$  converges to  $\nu^\ell$  for  $\ell = 1, \dots, L$ .

*Proof.* Consider the subsystem of  $M_j \in S_\ell$ , and  $c_i \in \mathcal{C}(S_\ell) \setminus \mathcal{C}(S_{\ell+1} \cup \dots \cup S_L)$ . With infinite supply of jobs of these types the system will be ergodic with stationary measure  $\nu^\ell$ . The Theorem again follows from the Lemma of Adan, Foss and Weiss [3].  $\square$

When there is no resource pooling and the arrival rate is high, the servers will split to subsets which move together at some joint average rate, tending to a stationary distribution of the queues inside each subset, but the queues separating these subsets of servers will grow without bound.

In general, fluid model equations (8), (9) do not determine the paths of  $\bar{Y}$ , and do not provide us with a way of verifying complete resource pooling or stability of the fluid model. This is not simply because we have not found the right fluid model equations necessary for that calculation. The fact is that for general bipartite graphs with service rates  $\mu_{s_j, c_i}$  that depend on both server and the customer type, under FCFS, first order and second order moment information alone does not determine the fluid limits of the system. This was discovered in the seminal paper of Foss and Chernova [15]. They consider a system with 3 servers, 3 customer types and an almost complete bipartite compatibility graph as illustrated in Figure 5. Here  $\alpha_{c_1} = \alpha_{c_2} = \alpha_{c_3} = 1/3$ , and

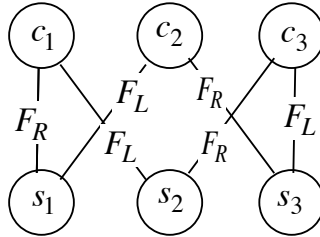


Figure 5: A symmetric system with an almost complete 3 server 3 customer types graph.

the service time distributions are  $v_{s_1, c_1} \sim F_R, v_{s_1, c_2} \sim F_L, v_{s_2, c_1} \sim F_L, v_{s_2, c_3} \sim F_R, v_{s_3, c_2} \sim F_R, v_{s_3, c_3} \sim F_L$ , with means  $m_L \neq m_R$ , so that each server has two service time distributions, and each customer type has two service time distributions. Foss and Chernova show that for some fixed  $\lambda, m_L, m_R$  it is possible to choose  $F_L, F_R$  in such a way that the system under FCFS (they actually consider the equivalent JSW policy) is positive Harris recurrent, but under a different choice of  $F_L, F_R$  it is transient.

In the rest of the paper we impose further assumptions on the service rates or on the shape of the bipartite graph, under which we derive more detailed fluid model equations. With the aid of these we can verify complete resource pooling and stability of the fluid limits, and find conditions under which they hold.

## 5 Service rates depend only on server

We now consider the special case where service rates depend only on the server (SD), and not on the customer type which he serves. We let  $m_{s_j}$  and  $\mu_{s_j} = 1/m_{s_j}$  be the mean service time and the service rate of server  $s_j$ . Define  $\mu = \sum_{s_j \in \mathcal{S}} \mu_{s_j}$  and  $\beta_{s_j} = \mu_{s_j}/\mu$ . Then  $\mu$  is the total service capacity of the system, and  $\beta_{s_j}$  is the fraction of service capacity provided by server  $s_j$ . For a subset  $S$  of server types we use the notation  $\mu_S = \sum_{s_j \in S} \mu_{s_j}$ ,  $\beta_S = \sum_{s_j \in S} \beta_{s_j}$ . In that case we have immediately:

**Corollary 2.** *Assume  $\mu_{s_j, c_i} = \mu_{s_j}$ ,  $c_i \in \mathcal{C}(s_j)$ ,  $j = 1, \dots, J$ . Under the conditions of Theorem 2 a.s. all fluid limits at  $s < \tau < t$  must satisfy:*

$$\dot{Y}_j(\tau) = \frac{\mu_{\{M_k, \dots, M_l\}}}{\alpha_{\mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)}}, \quad j = k, \dots, l. \quad (11)$$

*Proof.* Substituting  $\mu_{M_j, c_i} = \mu_{M_j}$  into (9), and summing over  $c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)$  we obtain:

$$\dot{Y}_l(\tau) \alpha_{\mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)} = \sum_{j=k}^l \mu_{M_j} \sum_{c_i \in \mathcal{C}(M_j) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)} \dot{T}_{M_j, c_i}(\tau)$$

and using (8) the corollary follows.  $\square$

This shows that in the SD special case, indeed all the fluid trajectories of  $\bar{Y}_j$  are along straight lines, as in Figure 4. The following theorems and definition characterize the fluid limits of  $\bar{Y}_j$  completely. The proofs of these theorems was given in Proposition B10 in [5]. We present a slightly simplified proof here for completeness.

**Condition for complete resource pooling in the SD case:** For every subset of servers  $S \neq \emptyset, \mathcal{S}$  and every subset of customer types  $C \neq \emptyset, \mathcal{C}$ , the following 3 equivalent conditions hold:

$$\beta_{S(C)} > \alpha_C, \quad \alpha_{\mathcal{C}(S)} > \beta_S, \quad \beta_S > \alpha_{\mathcal{U}(S)}. \quad (12)$$

The following Lemma has often been used in proofs of fluid stability (see [14]), and is useful here:

**Lemma 1.** *Let  $g(t)$  be an absolutely continuous nonnegative function on  $t \geq 0$  and let  $\dot{g}(t)$  denote its derivative whenever it exists.*

- (i) *If  $g(t) = 0$  and  $\dot{g}(t)$  exists, then  $\dot{g}(t) = 0$ .*
- (ii) *Assume that for some  $\epsilon > 0$ , whenever  $g(t) > 0$  and  $\dot{g}(t)$  exists, then  $\dot{g}(t) < -\epsilon$ . Then  $g(t) = 0$  for all  $t > \delta$  where  $\delta = g(0)/\epsilon$ . Furthermore  $g(\cdot)$  is nonincreasing and hence, once it reaches zero, it stays there forever.*

**Theorem 6.** (i) *Assume that condition (12) holds, then complete resource pooling holds, that is, for any initial conditions there exists  $t_0$  such that for every fluid limit  $\bar{Y}_1(t) = \dots = \bar{Y}_J(t) = \min(\mu t, \lambda t)$  holds for  $t > t_0$ .*

(ii) *Assume that condition (12) only holds with  $\geq$  instead of  $>$ . Then complete weak resource pooling holds.*

(iii) *Assume that complete resource pooling condition (12) is strictly violated. Then it is not possible to have  $\bar{Y}_1(\tau) = \dots = \bar{Y}_J(\tau) < \lambda \tau$  for all  $\tau$  in an interval  $s < \tau < t$ .*

*Proof.* (i) Assume that (12) holds, and that at time  $t$  the servers are split into the ordered partition  $\bar{\mathcal{S}}(t) = (\bar{S}_1, \dots, \bar{S}_L)$ , and each of these subsets of servers are moving together.

By Corollary 2,

$$\dot{Y}_1(t) = \mu \frac{\beta_{\bar{S}_1}}{\alpha_{\mathcal{U}(\bar{S}_1)}}, \quad \dot{Y}_J(t) = \min \left( \lambda, \mu \frac{\beta_{\bar{S}_L}}{\alpha_{\mathcal{C}(\bar{S}_L)}} \right).$$

By (12),  $\frac{\beta_{\bar{S}_1}}{\alpha_{\mathcal{U}(\bar{S}_1)}} > 1$  while  $\frac{\beta_{\bar{S}_L}}{\alpha_{\mathcal{C}(\bar{S}_L)}} < 1$ . Hence  $\dot{Y}_1(t) > \mu$  while  $\dot{Y}_J(t) \leq \min(\lambda, \mu)$ , so that  $\frac{d}{dt}(\bar{Y}_J(t) - \bar{Y}_1(t)) < 0$ .

By looking at the finite number of all different splits we can find  $\epsilon > 0$  such that  $\frac{d}{dt}(\bar{Y}_J(t) - \bar{Y}_1(t)) < \epsilon < 0$ . (i) then follows from Lemma 1.

(ii) Assume first that for some  $S$ ,  $\beta_S = \alpha_{\mathcal{U}(S)}$ , in which case also  $\beta_{\bar{S}} = \alpha_{\mathcal{C}(\bar{S})}$ , and consider the case that for all other subsets, (12) holds. Assume at time  $t$  a partition  $\bar{\mathcal{S}}(t) = (S_1, \dots, S_L)$  in which  $S_1$  is neither  $S$  nor

$\bar{S}$ . In that case by the argument of (i),  $\dot{Y}_1(t) > \dot{Y}_J(t)$ . This shows that for some  $t_0$  we have for all  $t > t_0$  the trajectories are given by the partition  $\bar{S}(t) = \{S, \bar{S}\}$ . The proof for any number of weak inequalities in (12) follows by induction.

(iii) If resource pooling is strictly violated then there exists a subset of the servers,  $S = \{M_1, \dots, M_L\}$ , such that  $\beta_S < \alpha_{\mathcal{U}(S)}$ . Assume contrary to the statement of the proposition that there exists a fluid limit for which  $\bar{Y}_1(\tau) = \dots = \bar{Y}_J(\tau) < \lambda\tau$  for  $\tau \in [t, t + \Delta]$ ,  $\Delta > 0$ . Denote the common value of  $\bar{Y}_j, \dot{Y}_j$ ,  $j = 1, \dots, J$  by  $\bar{Y}_{\text{Common}}, \dot{Y}_{\text{Common}}$ . Consider customer types  $c_i \in \mathcal{U}(M_1, \dots, M_L)$ . By (8)-(9) we have:

$$\dot{Y}_{\text{Common}}(\tau) \alpha_{\mathcal{U}(M_1, \dots, M_L)} = \sum_{j=1}^L \mu_{M_j} \left( \sum_{c_i \in \mathcal{U}(M_1, \dots, M_L)} \dot{T}_{M_j, c_i}(\tau) \right) \leq \sum_{j=1}^L \mu_{M_j}$$

Hence we obtain

$$\dot{Y}_{\text{Common}}(\tau) \leq \mu \frac{\beta_S}{\alpha_{\mathcal{U}(S)}} < \mu.$$

On the other hand, if  $\bar{Y}_1(\tau) = \dots = \bar{Y}_J(\tau) < \lambda\tau$  for  $\tau \in [t, t + \Delta]$ ,  $\Delta > 0$ , then by summing (9) over all servers and all customer types and using (8), we obtain  $\dot{Y}_{\text{Common}}(\tau) = \mu$ . This contradiction proves (ii).  $\square$

**Definition 3.** Consider a partition of the servers into subsets  $S', S, S''$ . We say that  $S$  satisfies complete resource pooling condition (12) between  $S'$  and  $S''$  (the order of  $S'$  before  $S''$  is important here), if the sub-system which consists of servers  $s_j \in S$ , and the customer types  $c_i \in \mathcal{C}(S) \setminus \mathcal{C}(S'')$ , with  $\tilde{\beta}_{s_j} = \beta_{s_j} / \beta_S$ ,  $\tilde{\alpha}_{c_i} = \alpha_{c_i} / \alpha_{\mathcal{C}(S) \setminus \mathcal{C}(S'')}$ , satisfies (12).

We now have the following theorem, which enables us to trace the exact piecewise linear trajectories of the fluid model of the system:

**Theorem 7.** Consider a fluid limit with  $\bar{Y}_{k-1}(t) < \bar{Y}_k(t) = \dots = \bar{Y}_l(t) < \bar{Y}_{l+1}(t) \leq \lambda t$  for some  $t$  and let  $\bar{S}(t) = (S', \{M_k, \dots, M_l\}, S'')$  be the corresponding partition of the servers. Then

$$\dot{Y}_j(t) = \mu \frac{\beta_{\{M_k, \dots, M_l\}}}{\alpha_{\mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(S'')}}, \quad j = k, \dots, l \quad (13)$$

during  $t < \tau < t + \Delta$  for some  $\Delta > 0$ , if and only if  $\{M_k, \dots, M_l\}$  satisfies complete resource pooling condition of Definition 3 between  $S'$  and  $S''$ .

*Proof.* If  $\bar{Y}_{k-1}(t) < \bar{Y}_k(t) = \dots = \bar{Y}_l(t) < \bar{Y}_{l+1}(t) \leq \lambda t$  then by continuity, for some  $\Delta$ :  $\bar{Y}_{k-1}(\tau) < \bar{Y}_k(\tau) \leq \dots \leq \bar{Y}_l(\tau) \leq \bar{Y}_{l+1}(\tau) \leq \lambda\tau$  for  $t < \tau < t + \Delta$ , and so going back to the originating  $\omega$  and subsequence  $r$  for large enough  $r$ , we will have  $\bar{Y}_{k-1}^r(r\tau) < \bar{Y}_k^r(r\tau) \leq \dots \leq \bar{Y}_l^r(r\tau) \leq \bar{Y}_{l+1}^r(r\tau) \leq \lambda r\tau$  for  $rt < r\tau < rt + r\Delta$ . In other words, servers  $M_k, \dots, M_l$  will serve customer types  $c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)$  as an isolated FCFS-ALIS sub-system, in the time interval  $(rt, rt + r\Delta)$ . The theorem then follows by applying Theorem 6 to this subsystem.  $\square$

**Corollary 3.** Under complete resource pooling, the fluid model is stable if and only if  $\lambda < \mu$

It is shown in [5] that if resource pooling does not hold then there exists a unique decomposition of the system into subsystems  $(\mathcal{S}^{(1)}, \mathcal{C}^{(1)}), \dots, (\mathcal{S}^{(L)}, \mathcal{C}^{(L)})$  with  $\mathcal{C}^{(\ell)} = \mathcal{U}(\mathcal{S}_1 \cup \dots \cup \mathcal{S}^{(\ell)}) \setminus \mathcal{U}(\mathcal{S}_1 \cup \dots \cup \mathcal{S}^{(\ell-1)})$  and service rates  $\mu^{(\ell)} = \mu \beta_{\mathcal{S}^{(\ell)}}$ , and there are then values  $\lambda^{(1)} < \dots < \lambda^{(L)}$  so that system  $(\mathcal{S}^{(\ell)}, \mathcal{C}^{(\ell)})$  on its own is stable for all  $\lambda < \lambda^{(\ell)}$ , and the combined system exhibits local stability. These results carry over to our system.

In summary, for the SD case we get the complete traces of the fluid model of the system, including answers to questions of stability, resource pooling, or decomposition, under FCFS policy. In fact the fluid models are independent of the service time distributions, and depend only on first order moments. In particular, the results are the same as those obtained for the system with Poisson arrivals and exponential service rates.

On the question of matching rates, the fluid model is not informative enough. While we can obtain matching rates in the Poisson-exponential case as done in [4], we cannot calculate matching rates for general service time distributions in the SD case. We return to this question in Section 9. Matching rates can be calculated for some special bipartite graphs — we do that in Section 7.

## 6 Service rates depend only on customer type

We now consider the special case where service rates depend only on the customer type (CD), regardless of which of the compatible servers is serving. We let  $m_{c_i}$  and  $\mu_{c_i} = 1/m_{c_i}$  be the mean service time and the service rate for customer type  $c_i$ . In that case the total service capacity of the system is  $|\mathcal{S}| = J$ , for the  $J$  servers, but capacity for each subset  $C$  of customer types is  $|\mathcal{S}(C)|$ , the number of compatible servers.

In that case we have immediately:

**Corollary 4.** *Assume  $\mu_{s_j, c_i} = \mu_{c_i}$ , for  $s_j \in \mathcal{S}(c_i)$ ,  $i = 1, \dots, I$ . Under the conditions of Theorem 2 a.s. all fluid limits at  $s < \tau < t$  must satisfy:*

$$\dot{Y}_j(\tau) = \frac{l - k + 1}{\sum_{c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)} \alpha_{c_i} m_{c_i}}, \quad j = k, \dots, l. \quad (14)$$

*Proof.* Substituting  $\mu_{M_j, c_i} = \mu_{c_i}$  into (9), we have:

$$\dot{Y}_j(\tau) \alpha_{c_i} m_{c_i} = \sum_{j=k}^l \dot{T}_{M_j, c_i}(\tau), \quad c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J).$$

and summing over all  $c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)$  we get by (8) that

$$\dot{Y}_j(\tau) \sum_{c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(M_{l+1}, \dots, M_J)} \alpha_{c_i} m_{c_i} = l - k + 1.$$

□

This shows that also in the special case of CD all the fluid trajectories of  $\bar{Y}_j$  are along straight lines, as in Figure 4. The following definition and theorems characterize the fluid limits of  $\bar{Y}_j$  completely.

**Condition for complete resource pooling in the CD case:** For every subset of servers  $C \neq \emptyset, \mathcal{C}$ :

$$\frac{|\mathcal{S}(C)|}{|\mathcal{S}|} > \frac{\sum_{c_i \in C} \alpha_{c_i} m_{c_i}}{\sum_{c_i \in \mathcal{C}} \alpha_{c_i} m_{c_i}}. \quad (15)$$

**Theorem 8.** (i) *If condition (15) holds then complete resource pooling holds, i.e., for some  $t_0$  and for any  $\lambda$ ,  $\bar{Y}_1(t) = \dots = \bar{Y}_J(t)$  for all  $t > t_0$ .*

(ii) *If (15) holds only with  $\geq$  replacing  $>$ , then complete weak resource pooling holds.*

(iii) *If (15) is strictly violated then complete resource pooling does not hold.*

*Proof.* If  $\bar{Y}_1(\tau) = \dots = \bar{Y}_J(\tau)$  for  $s < \tau < t$  then by (14)

$$\dot{Y}_j(\tau) = \frac{|\mathcal{S}|}{\sum_{c_i \in \mathcal{C}} \alpha_{c_i} m_{c_i}} \quad (16)$$

Assume that  $\bar{Y}_J(t) > \bar{Y}_1(t)$  and assume partition  $S_1(t), \dots, S_L(t)$ . We show that if (15) holds then there exists  $\varepsilon > 0$  such that  $\dot{Y}_1(t) - \dot{Y}_J(t) \geq \varepsilon$ , which by Lemma 1 proves that complete resource pooling holds.

Indeed, by Corollary 4 and (15):

$$\dot{Y}_{S_1} = \frac{|S_1|}{\sum_{c_i \in \mathcal{U}(S_1)} \alpha_{c_i} m_{c_i}} > \frac{|\mathcal{S}|}{\sum_{c_i \in \mathcal{C}} \alpha_{c_i} m_{c_i}},$$

On the other hand:

$$\frac{|\mathcal{S}|}{\sum_{c_i \in \mathcal{C}} \alpha_{c_i} m_{c_i}} = \frac{|\mathcal{S} \setminus S_L| + |S_L|}{\sum_{c_i \in \mathcal{C} \setminus \mathcal{C}(S_L)} \alpha_{c_i} m_{c_i} + \sum_{c_i \in \mathcal{C}(S_L)} \alpha_{c_i} m_{c_i}}$$

and  $\mathcal{C} \setminus \mathcal{C}(S_L) = \mathcal{C}(\mathcal{S} \setminus S_L)$ , and hence  $\frac{|\mathcal{S} \setminus S_L|}{\sum_{c_i \in \mathcal{C} \setminus \mathcal{C}(S_L)} \alpha_{c_i} m_{c_i}} > \frac{|\mathcal{S}|}{\sum_{c_i \in \mathcal{C}} \alpha_{c_i} m_{c_i}}$  which implies that:

$$\dot{Y}_{S_L} = \frac{|S_L|}{\sum_{c_i \in \mathcal{C}(S_L)} \alpha_{c_i} m_{c_i}} < \frac{|\mathcal{S}|}{\sum_{c_i \in \mathcal{C}} \alpha_{c_i} m_{c_i}},$$

The proof of (ii) is similar to the proof of (ii) in Theorem 6.

If the condition (15) is strictly violated then clearly it is not possible to have  $\bar{Y}_1(\tau) = \dots = \bar{Y}_J(\tau)$  for  $s < \tau < t$ . If it is only weakly violated, i.e., there exists  $C, \mathcal{S}(C)$  such that  $\frac{|\mathcal{S}(C)|}{|\mathcal{S}|} = \frac{\sum_{c_i \in C} \alpha_{c_i} m_{c_i}}{\sum_{c_i \in \mathcal{C}} \alpha_{c_i} m_{c_i}}$ , then if initially servers  $\mathcal{S}(C)$  are behind all the others, they will never catch up with  $\bar{Y}_J$ .  $\square$

All the trajectories of the fluid limits for the CD case can be determined by the following Corollary, which mimics Theorem 7, and has the same proof.

**Corollary 5.** *Consider a fluid limit with  $\bar{Y}_{k-1}(t) < \bar{Y}_k(t) = \dots = \bar{Y}_l(t) < \bar{Y}_{l+1}(t) \leq \lambda t$  for some  $k, l$ , and  $t$ , and let  $\bar{\mathcal{S}}(t) = (S', \{M_k, \dots, M_l\}, S'')$  be the corresponding partition of the servers. Then*

$$\dot{\bar{Y}}_j(t) = \frac{l - k + 1}{\sum_{c_i \in \mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(S'')} \alpha_{c_i} m_{c_i} / \alpha_{\mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(S'')}} \quad (17)$$

during  $t < \tau < t + \Delta$  for some  $\Delta > 0$ , if and only if the subsystem consisting of  $\{M_k, \dots, M_l\}$  and  $\mathcal{C}(M_k, \dots, M_l) \setminus \mathcal{C}(S'')$  satisfies condition (15).

As in the SD case, we get a complete picture of the fluid model in the CD case. Similar to the SD case however, the fluid model does not contain enough information to calculate the matching rates.

## 7 Systems with computable matching rates

For some types of compatibility graphs it is possible to calculate the matching rates of the fluid model, and in those cases one can again show that the fluid levels are piecewise linear, and calculate their trajectories. We consider two such special types of networks: networks with complete bipartite compatibility graph and networks with tree compatibility graph, as well as their hybrid. The fluid models for these systems under FCFS were considered by Talreja and Whitt [28], for the SD case. In our derivations here we allow service rates to depend both on the server and on the customer type.

### 7.1 Network with complete bipartite compatibility graph

We now assume that every server can serve all types of customers, i.e., the compatibility graph is a complete bipartite graph. If all the servers can serve all the customers, then servers will never skip customers, and in effect the system will just behave like a GI/GI/J queueing system with non-identical servers.

When a server will complete service he will immediately overtake all the other servers and will start serving the first waiting customer. Average service time for server  $s_j$ , service rate for server  $s_j$ , and total service capacity of the system, are then:

$$\mu = \sum_{j=1}^J \mu_{s_j}, \quad \mu_{s_j} = m_{s_j}^{-1}, \quad m_{s_j} = \sum_{c_i \in \mathcal{C}(s_j)} \alpha_{c_i} m_{s_j, c_i}, \quad (18)$$

and we can calculate matching rates as follows:

$$r_{s_j, c_i} = \frac{\mu_{s_j}}{\mu} \frac{\alpha_{c_i}}{\sum_{c_k \in \mathcal{C}(s_j)} \alpha_{c_k}}. \quad (19)$$

Using the same arguments as for GI/GI/J we get:

**Theorem 9.** *For the case of a complete bipartite compatibility graph, under FCFS-ALIS policy, there is complete resource pooling always, and for every fluid model almost surely*

$$\bar{Y}_1(t) = \dots = \bar{Y}_J(t) = \min(\bar{Y}_J(0) + \mu t, \lambda t), \quad t > 0,$$

where  $\mu$  is given in (18). The matching rates while  $\bar{Y}_J(t) < \lambda t$  are given by  $r_{s_j, c_i}$  in (19).

*Proof.* Recall that  $\bar{Y}_J(0) \leq 0$  in our system description. The matching rates correspond to the fact that server devotes a fraction  $\alpha_{c_i} / \sum_{c_k \in \mathcal{C}(s_j)} \alpha_{c_k}$  of his services to type  $c_i$ . The service rate of each server is given by (18) as long as there is a queue, and so his fraction of all services is  $\mu_{s_j} / \mu$ , and (19) follows.  $\square$



## 7.2 Network with tree bipartite compatibility graph

A tree graph is a connected graph with no loops. With  $K$  nodes it will have exactly  $K - 1$  edges, and it will always have at least two leaves (nodes that are connected by a single edge). Furthermore, any sub-graph will be a tree, or a union of disconnected trees (a forest). We now assume that the bipartite graph  $\mathcal{G}$  is a tree. It has  $I + J$  nodes and therefore it has  $I + J - 1$  compatible pairs (edges), and at least two leaves, each of which can be either a server or customer type.

Let  $S_1, \dots, S_L$  be an ordered partition of  $s_1, \dots, s_J$ . Denote by  $C_\ell = \mathcal{C}(S_\ell) \setminus \mathcal{C}(S_{\ell+1} \cup \dots \cup S_L)$ . Consider now a fluid limit  $\bar{Y}_1(\tau), \dots, \bar{Y}_J(\tau)$ , with permutation  $\bar{M}_1(\tau), \dots, \bar{M}_J(\tau)$ , and assume that for all  $s < \tau < t$  the following holds:

$$\begin{aligned} \bar{M}_j(\tau), \bar{M}_{j'}(\tau) \in S_\ell &\implies \bar{P}_{M_j}(\tau) = \bar{P}_{M_{j'}}(\tau), \\ \bar{M}_j(\tau) \in S_\ell \text{ and } \bar{M}_{j'}(\tau) \in S_{\ell+1} &\implies \bar{P}_{M_j}(\tau) < \bar{P}_{M_{j'}}(\tau) \text{ or } \bar{P}_{M_j}(\tau) = \bar{P}_{M_{j'}}(\tau) \end{aligned} \quad (20)$$

and the subgraphs of  $(S_\ell, C_\ell)$ ,  $(S_{\ell+1}, C_{\ell+1})$  are not connected

We denote the common value of  $\bar{P}_{M_j}(\tau)$  for  $M_j \in S_\ell$  by  $\bar{Y}_{S_\ell}(\tau)$ . Clearly, by the continuity of the  $\bar{Y}_j(\cdot)$ , such a partition is defined for every  $t$  and for some  $s < t$ . This partition is a refinement of the partitions discussed in Sections 3, 5, where we further divide subset of servers that move together, so that each such subset will have the property that each  $(S_\ell, C_\ell)$  is connected.

**Theorem 10.** *Assume that the bipartite compatibility graph is a tree, and consider the partition  $(S_1, \dots, S_L)$  as in (20) valid for  $s < \tau < t$ . Then:*

(i) *Equations (8)-(9) have a unique solution, for every  $S_\ell \in (S_1, \dots, S_L)$ , and hence  $\dot{\bar{T}}_{M_j, c_i}(\tau), \dot{\bar{Y}}_j(\tau)$  are constant for  $s < \tau < t$ . As a result, almost surely, the fluid limit has unique continuous piecewise linear trajectories.*

(ii) *Consider the set of equations*

$$\begin{aligned} \sum_{c_i \in \mathcal{C}(s_j)} \eta_{s_j, c_i} &= 1 \quad j = 1, \dots, J, \\ \sum_{s_j \in \mathcal{S}(c_i)} \frac{\mu_{s_j, c_i}}{\alpha_{c_i}} \eta_{s_j, c_i} &= \mu, \quad i = 1, \dots, I \end{aligned} \quad (21)$$

*with the  $I + J - 1$  unknowns  $\eta_{s_j, c_i}$ ,  $(s_j, c_i) \in \mathcal{G}$ , and an additional unknown  $\mu$ . The system will have complete resource pooling if and only if (21) has a positive solution, and it will have complete weak resource pooling if the solution is non-negative.*

(iii) *If complete resource pooling holds then  $\mu$  is the pooled service rate, and the matching rates are given by:*

$$r_{s_j, c_i} = \frac{\mu_{s_j, c_i} \eta_{s_j, c_i}}{\mu}. \quad (22)$$

*Proof.* (i) The equations (8)-(9) for each  $S_\ell$  are:

$$\begin{aligned} \sum_{c_i \in C_\ell} \dot{\bar{T}}_{M_j, c_i}(\tau) &= 1, \quad M_j \in S_\ell, \\ \dot{\bar{Y}}_{S_\ell} &= \sum_{s_j \in S_\ell} \frac{\mu_{M_j, c_i}}{\alpha_{c_i}} \dot{\bar{T}}_{M_j, c_i}(\tau), \quad c_i \in C_\ell. \end{aligned}$$

with unknowns  $\dot{\bar{Y}}_{S_\ell}$  and  $\dot{\bar{T}}_{M_j, c_i}(\tau)$  for each edge in the subgraph  $(S_\ell, C_\ell)$ . Since the subgraph is a connected tree, the number of unknowns is equal to the number of equations. The equations are independent, so the solution is unique, for all  $t < \tau < s$ . The solution must be non-negative, because the fluid limits exist. This proves that the fluid limit  $\bar{Y}_{S_\ell}$  moves along a linear trajectory in the interval  $(s, t)$ .

We note that the equations can be solved in  $|S_\ell| + |C_\ell|$  steps: Locate a leaf in the graph. If it is  $M_j$ , it has a single customer type  $c_i = \mathcal{C}(M_j) \cap C_\ell$ , and  $\dot{\bar{T}}_{M_j, c_i}(\tau) = 1$ . If it is  $c_i$  it has a single server  $M_j = \mathcal{S}(c_i) \cap S_\ell$ , and then  $\dot{\bar{T}}_{M_j, c_i} = \dot{\bar{Y}}_{S_\ell} \alpha_{c_i} m_{M_j, c_i}$ . In either case one can eliminate the leaf node and one equation and continue to solve for the remaining graph. Note that deleting a leaf from a tree leaves a connected tree.

(ii) Clearly if there is no positive solution to (21) then there can be no complete resource pooling (i.e., it is impossible to have  $\bar{Y}_1(\tau) = \dots = \bar{Y}_J(\tau)$  for any interval of  $\tau$ 's). If the solution is non-negative with some 0 values for some edges, this implies that some disconnected subtrees move at the same rate, but may have different initial positions. So the system has complete weak resource pooling. Assume that (21) has a positive solution. We need to show that for some  $t_0$  all the trajectories  $\bar{Y}_j(t)$  meet for  $t > t_0$ . Assume that at time  $t$ ,  $\bar{Y}_1(t) < \bar{Y}_J(t)$ . We will show that  $\dot{\bar{Y}}_J(t) - \dot{\bar{Y}}_1(t) < 0$ , which by Lemma 1 will complete the proof.

By continuity we have for an interval  $t - \delta < \tau < t + \delta$  in which the partition is  $S = (S_1, \dots, S_L)$ , where  $S_1 = (M_1, \dots, M_k)$  and  $S_L = (M_l, \dots, M_J)$  and  $\bar{Y}_{S_1}(\tau) < \bar{Y}_{S_L}(\tau)$ . We will show that  $\dot{\bar{Y}}_{S_1}(\tau) > \dot{\bar{Y}}_{S_L}(\tau)$ .

Denote by  $\mu^{(S_1)}, \eta_{M_j, c_i}^{(S_1)}$  and  $\mu^{(S_L)}, \eta_{M_j, c_i}^{(S_L)}$  the non-negative solutions of (8)-(9) for  $S_1$  and for  $S_L$ , and by  $\mu^{(0)}, \eta_{s_j, c_i}^{(0)}$  the positive solution of (21).

We note that the solution of (8)-(9) for the tree graphs  $(S_\ell, C_\ell)$ ,  $\ell = 1, \dots, L$ , as well as for the complete tree graph  $(S, C)$  are in fact the unique optimal solutions of the corresponding linear programs (LP):

$$\begin{aligned} & \max \mu \\ & \text{s.t.} \begin{cases} \sum_{c_i \in C_\ell} \eta_{M_j, c_i} \leq 1, & M_j \in S_\ell, \\ \sum_{M_j \in S_\ell} \mu_{M_j, c_i} \eta_{M_j, c_i} \geq \mu \alpha_{c_i}, & c_i \in C_\ell, \\ \eta_{M_j, c_i} \geq 0. \end{cases} \end{aligned}$$

The fact that they are unique optimal solutions is explained in the following Section 8.

Consider then the LP (23) for  $(S_1, C_1)$ , and substitute the values of  $\mu^{(0)}, \eta_{M_j, c_i}^{(0)}$ . We then have that:

$$\sum_{M_j \in S_1} \mu_{M_j, c_i} \eta_{M_j, c_i}^{(0)} = \mu^{(0)} \alpha_{c_i}, \quad c_i \in C_1,$$

because  $\mathcal{S}(C_1) = S_1$ , since  $C_1$  includes customers that were skipped by all the other servers. At the same time:

$$\sum_{c_i \in C_\ell} \eta_{M_j, c_i}^{(0)} < 1, \quad \text{for at least one } M_j \in S_1,$$

because the graph of  $(S, C)$  is connected, and therefore there exists a link from some server  $M_j \in S_1$  to a customer type  $c_i \notin C_1$ , and by assumption  $\eta_{M_j, c_i}^{(0)} > 0$ . Hence this is a feasible but not optimal solution, which proves that  $\mu^{(S_1)} > \mu^{(0)}$ .

On the other hand, consider the LP (23) for  $(S_L, C_L)$ . Because  $C_L = C \cap C(S_L)$ , it has all the constraints as the LP for  $(S, C)$ , with the additional constraints that  $\eta_{M_j, c_i} = 0$  whenever  $M_j \notin S_L$ . Hence the LP for  $(S_L, C_L)$  is more constrained than that for  $S, C$ , and further more, in the optimal solution of  $S, C$  all the  $\eta_{M_j, c_i}^{(0)} > 0$ . This implies that  $\mu^{(S_L)} < \mu^{(0)}$ .

But,  $\mu^{(S_1)} = \dot{\bar{Y}}_{S_1} = \dot{\bar{Y}}_1$ ,  $\mu^{(S_L)} = \dot{\bar{Y}}_{S_L} = \dot{\bar{Y}}_J$ , and we have shown that if  $\bar{Y}_1(t) < \bar{Y}_J(t)$  then  $\dot{\bar{Y}}_1(t) - \dot{\bar{Y}}_J(t) > 0$ , as required.

(iii) In the optimal solution of (21) the values of  $\eta_{s_j, c_i}$  are the fractions of time allocated by server  $s_j$  to customers of type  $c_i$ , and therefore the rate at which customers of type  $c_i$  are processed by server  $s_j$  is  $\mu_{s_j, c_i} \eta_{s_j, c_i}$ . The total processing rate is then the sum of all these  $\mu = \sum_{s_j, c_i \in \mathcal{G}} \mu_{s_j, c_i} \eta_{s_j, c_i}$ , which is indeed the solution of (21). The matching rates are therefore given by (22).  $\square$

**Remark 1.** A system is a hybrid of the systems studied in Sections 7.1-7.2, if its bipartite compatibility graph consists of several complete graphs which are connected by a tree graph. For these hybrid systems one can again calculate the matching rates, and obtain a complete description of the fluid model trajectories.

## 8 Maximal throughput under FCFS

We consider a static planning problem similar to Harrison and Lopez [22]:

$$\begin{aligned} & \max \mu \\ & \text{s.t.} \begin{cases} \sum_{c_i \in \mathcal{C}(s_j)} \eta_{s_j, c_i} \leq 1, & s_j \in \mathcal{S}, \\ \sum_{s_j \in \mathcal{S}(c_i)} \mu_{s_j, c_i} \eta_{s_j, c_i} \geq \alpha_{c_i} \mu, & c_i \in \mathcal{C}, \\ \eta_{s_j, c_i} \geq 0, & (s_j, c_i) \in \mathcal{G} \end{cases} \end{aligned} \quad (23)$$

with the decision variables  $\eta_{s_j, c_i}$ ,  $(s_j, c_i) \in \mathcal{G}$  and  $\mu$ . Here  $\eta_{s_j, c_i}$  is the fraction of time that server  $s_j$  allocates to customers of type  $c_i$ , and  $\mu$  is the rate at which the total stream of arrivals is served. The first  $J$  constraints (the server constraints) say that the sum of allocations for each server cannot exceed 1. The next  $I$  constraints (the customer constraints) say that the allocations  $\eta_{s_j, c_i}$  are sufficient to serve the fraction customers of type  $c_i$ , to keep up with the total service rate  $\mu$ . In terms of our system,  $\eta_{s_j, c_i}$  can be thought of as long term average of  $\hat{T}_{s_j, c_i}$ , and  $\mu$  can be thought of as long term average of  $\bar{Y}_1$ , the rate of progress of  $\bar{Y}_1$ .

The following Theorem is a simple consequence of Theorem 1 in the paper of Dai and Lin [13]

**Theorem 11** (Dai-Lin [13]). *Let  $\mu^*$  be the optimal value of the LP (23). Then under any policy, the fluid model is unstable if  $\lambda > \mu^*$ , so any policy that achieves fluid stability for all  $\lambda < \mu^*$  is throughput optimal.*

*Proof.* Consider the departure processes of customers of type  $c_i$ . Denote its fluid limits by  $\bar{D}_{c_i}(t)$ , and let the fluid allocation rates be  $\bar{T}_{s_j, c_i}(t)$ . Under any policy,  $\sum_{c_i \in \mathcal{C}(s_j)} \bar{T}_{s_j, c_i}(t) \leq 1$  needs to hold for all  $t > 0$  for all servers. Also, for any fluid limit, under any policy  $\bar{D}_{c_i}(t) = \sum \mu_{s_j, c_i} \bar{T}_{s_j, c_i}(t)$ . It follows that the fluid model can only be stable if for every  $c_i$ ,  $\bar{D}_{c_i}(t) = \sum \mu_{s_j, c_i} \bar{T}_{s_j, c_i}(t) \geq \lambda \alpha_{c_i}$ . Hence  $\mu^*$  is an upper bound on  $\lambda$  for which the fluid model can be stable.  $\square$

We note that (23) is an optimization problem for a network with gains (cf. [6]). We now proceed to discuss the solution of the problem (23) through a number of observations.

- (i) The problem is feasible, since 0 for all decision variables is a solution.
- (ii) The problem is bounded, since  $\mu$  is bounded by a positive linear combination of the  $\eta_{s_j, c_i}$ , and each  $\eta_{s_j, c_i} \leq 1$ .
- (iii) The optimal value is  $\mu^* > 0$ , since the problem is feasible if we take  $\eta_{s_j, c_i} = \frac{1}{IJ}$ .
- (iv) The server constraints are satisfied as equalities in the optimal solution, since  $\mu$  can only increase with every  $\eta_{s_j, c_i}$ .

We rewrite the LP and its dual, DP, in a slightly different form, including slack variables:

$$\begin{aligned} \text{LP} \quad & \begin{cases} \max \mu \\ \text{s.t.} \begin{cases} \sum_{c_i \in \mathcal{C}(s_j)} \eta_{s_j, c_i} = 1, & s_j \in \mathcal{S}, \\ \mu - \sum_{s_j \in \mathcal{S}(c_i)} \frac{\mu_{s_j, c_i}}{\alpha_{c_i}} \eta_{s_j, c_i} + \theta_{c_i} = 0, & c_i \in \mathcal{C}, \\ \eta_{s_j, c_i}, \theta_{c_i} \geq 0, & (s_j, c_i) \in \mathcal{G}. \end{cases} \end{cases} \\ \text{DP} \quad & \begin{cases} \min \sum_{s_j \in \mathcal{S}} y_{s_j} \\ \text{s.t.} \begin{cases} \sum_{c_i \in \mathcal{C}} z_{c_i} = 1, \\ y_{s_j} - \frac{\mu_{s_j, c_i}}{\alpha_{c_i}} z_{c_i} - x_{s_j, c_i} = 0, & (s_j, c_i) \in \mathcal{G}, \\ z_{c_i}, x_{s_j, c_i} \geq 0, & (s_j, c_i) \in \mathcal{G}. \end{cases} \end{cases} \end{aligned}$$

We observe that:

- (v) In the optimal solution there is at least one  $\eta_{s_j, c_i} > 0$  for each server  $s_j$ , and at least one  $\eta_{s_j, c_i} > 0$  for each customer type  $c_i$ , since the server constraints are satisfied as equalities, and in the customer constraints  $\mu > 0$ .
- (vi) Every basic optimal solution has no less than  $\min\{I, J\}$  and no more than  $I + J - 1$  positive  $\eta_{s_j, c_i}$ , by (v) and since there are  $I + J$  constraints and  $\mu > 0$ .
- (vii) Since the primal is feasible and bounded, both the primal and the dual possess optimal solutions.
- (viii) In an optimal solution  $y_{s_j} \geq 0$ , since it needs to be  $\geq$  than non-negative quantities.

The most important property of the solutions is the following results, which must be known and hidden in the literature on network flows with gains, but we could not find a good explicit reference and we provide a proof here.

**Lemma 2.** *The positive arcs in a basic solution of the LP (23) cannot contain a cycle.*

*Proof.* Assume that a basic solution of LP (23) contains the columns of a cycle of arcs of  $\mathcal{G}$ , which for simplicity we assume are labeled as  $(s_1, c_1), (s_1, c_2), (s_2, c_2), (s_2, c_3), \dots, (s_{L-1}, c_L), (s_L, c_L), (s_L, c_1)$ . We will get a contradiction. Denote  $a_{j,i} = \frac{\mu_{s_j, c_i}}{\alpha_{c_i}}$ , for these arcs. Consider the complementary slack dual solution. It will have  $x_{s_i, c_i} = 0$  for all the  $2L$  arcs in the cycle. That implies that  $y_{s_j} = a_{j,i} z_{c_i}$  for all these arcs. This implies that  $\frac{a_{1,1} a_{2,2} \dots a_{L,L}}{a_{1,2} a_{2,2} \dots a_{L,1}} = 1$ . Consider now the square matrix formed by  $2L$  columns corresponding to these arcs, and the  $2L$  non-zero rows of these columns. Its determinant is  $a_{1,1} a_{2,2} \dots a_{L,L} - a_{1,2} a_{2,2} \dots a_{L,1} = 0$ . Hence this cannot be a basis.  $\square$

In an optimal solution of the LP (23) we refer to the arcs with positive values of  $\eta_{s_j, c_i}$  as the solution graph.

**Theorem 12.** *Consider an optimal basic solution of the LP (23).*

(i) *Assume all the slacks  $\theta_{c_i} = 0$ , the optimal solution graph is a tree, and all the  $I + J - 1$  basic variables  $\eta_{s_j, c_i} > 0$ . Then erasing all the non-basic arcs and using FCFS for the remaining graph will achieve complete resource pooling and be throughput optimal with processing rate  $\mu^*$ .*

(ii) *Assume all the slacks  $\theta_{c_i} = 0$ , the optimal solution graph is a tree, but some of the  $I + J - 1$  basic variables are  $= 0$ . Then erasing all the non-basic arcs, and all the basic arcs with  $\eta_{s_j, c_i} = 0$ , and using FCFS for the remaining graph will achieve complete weak resource pooling and be throughput optimal with processing rate  $\mu^*$ .*

(iii) *If for some  $c_i$ ,  $\theta_{c_i} > 0$ , let  $C_1 = \{c_i : \theta_{c_i} = 0\}$  and let  $S_1 = \mathcal{S}(C_1)$ , and assume that the subgraph  $S_1, C_1$  is connected. Then in the solution graph  $S_1, C_1$  are not connected to the remaining nodes. Furthermore: formulate the LP (23) for the subsystem  $S_1, C_1$  with the corresponding arcs of  $\mathcal{G}$ . Then for this smaller problem either (i) or (ii) holds, and under FCFS complete resource pooling holds, the processing rate is  $\mu_1 = \mu^*$  and this policy is throughput optimal for  $C_1$ .*

(iv) *In the case of (iii), formulating the LP (23) for the subgraph of  $\mathcal{S} \setminus S_1, \mathcal{C} \setminus C_1$ , the optimal solution will have  $\mu_2 > \mu_1$ . Continuing in this way one obtains a unique decomposition of the the system to subgraphs  $(S_1, C_1), \dots, (S_L, C_L)$  each of which has an optimal tree solution, such that under FCFS it will have complete resource pooling, moving at rates  $\mu_L > \dots > \mu_1$ , and these rates are maximal throughput for  $(S_\ell, C_\ell)$  conditional on retaining the solutions of  $(S_1, C_1), \dots, (S_{\ell-1}, C_{\ell-1})$ .*

*Proof.* (i) If in an optimal basic solution all the slacks  $\theta_{c_i} = 0$  then the solution will have  $\mu > 0$  and  $I + J - 1$  basic variables  $\eta_{s_j, c_i}$  which by Lemma 2 have no loops and hence the solution graph is a tree. We assume that all the arcs in the tree have  $\eta_{s_j, c_i} > 0$ . If we use only the arcs of the tree, we have a system with a tree bipartite graph, and by Section 7.2, this system under FCFS will have complete resource pooling and processing capacity  $\mu^*$ . By Theorem 11 this will be throughput optimal.

(ii) If some of the arcs of the solution tree have  $\eta_{s_j, c_i} = 0$ , then by Theorem 10 the system with only the arcs of the solution graph under FCFS will have complete weak resource pooling, with processing capacity  $\mu^*$ . By Theorem 11 this will be throughput optimal

(iii) Consider  $c_i \in C_1$ , and assume that for some  $s_j \in \mathcal{S}(c_i)$  and  $c_k \notin C_1$ , the optimal solution has  $\eta_{s_j, c_k} > 0$ . it is then possible to reduce  $\eta_{s_j, c_k} > 0$  and increase  $\eta_{s_j, c_i}$  for all  $c_i \in C_1$ , without violating the feasibility of the solution. But this modified solution can only increase the objective value. This proves that in the solution graph  $S_1, C_1$  is not connected to any other parts of the system. Hence, solving the reduced problem for  $S_1, C_1$  the optimal solution will have  $\theta_{c_i} = 0$ ,  $c_i \in C_1$ , the solution graph will be a tree, and the optimal value for the reduced problem will be  $\mu_1 = \mu^*$ .

(iv) Clearly if for some  $c_i$ ,  $\theta_{c_i} > 0$  then there must exist  $(C_1, S_1)$  with a connected subgraph such that the conditions of (c) hold and  $\mu_1$  equal to the optimal  $\mu^*$  the value for the whole system. This then is maximum throughput for  $C_1$ . If we remove this  $C_1$  and its servers, we can continue to decompose the remaining graph.  $\square$

The results of Theorem 12 are for a particular basic solution. If there are several basic solutions, one might ask whether when using all the arcs of a non-basic solution and FCFS policy, there will be complete resource pooling and maximum throughput. We do not currently know the answer in general. For the special case of customer dependent service (CD) the answer is positive: As shown in Section 6, using the full bipartite graph we get complete resource pooling and maximum throughput under condition (15).

## 9 Exploration of the server dependent case under general service distributions

We have shown in Section 5 that the fluid model for the server-dependent case with general renewal arrivals and service times is the same as the fluid model for the Poisson exponential case. In particular, the necessary and sufficient condition for complete resource pooling is insensitive to the service time distribution. However, our fluid model analysis does not provide enough information to calculate the matching rates  $r_{ij}$  for this more general case. It is tempting to conjecture that if the system is overloaded, the matching rates will be the same as for the Poisson exponential case, and thus will be given by the matching rates calculated for the FCFS infinite bipartite matching model of [4].

This, however, is not the case. A simulation study reveals that in general, the matching rates in an overloaded, server-dependent system are sensitive to the service time distribution. The matching rates of such a system under non-exponential service time distribution are very close to those under the exponential distribution, yet they are different, in a statistically significant manner.

We considered three topologies for the system, labeled 1–3, shown in Figure 6. The topologies were parameterized by the customer type probabilities  $\alpha$  and service rates  $\mu$  as shown in Table 1. For each topology, we used three service time distributions: Pareto (denoted by the subscript ‘p’), and two versions of the uniform distribution (‘u<sub>1</sub>’ and ‘u<sub>2</sub>’). In the Pareto case, we used a distribution having the density  $f(x) = 3\gamma(\gamma x + 1)^{-4}$ ,  $x \geq 0$ . This Pareto distribution has only first and second finite moments, and is parameterized by a scale parameter  $\gamma$ , so that its mean is  $1/2\gamma$ . Thus, to achieve a service rate  $\mu_j$ , we set  $\gamma = \mu_j/2$ . The two uniform distributions are  $U(0, 2/\mu_j)$  and  $U(.9/\mu_j, 1.1/\mu_j)$ .

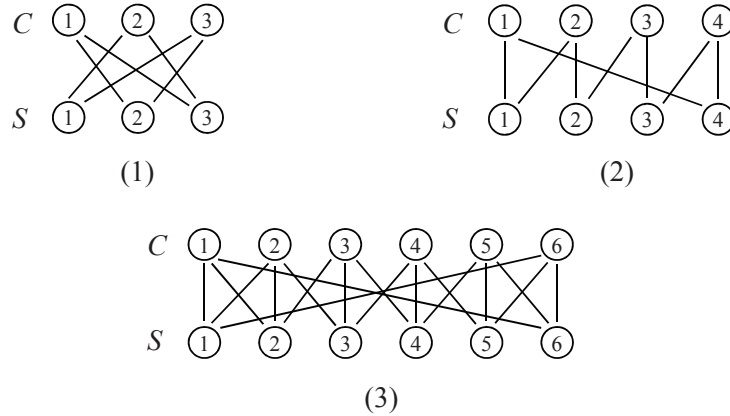


Figure 6: Topologies of the systems for the simulation study.

system	$\alpha$	$\mu$	exponential	Pareto	Uniform 1	Uniform 2
1	(.2, .6, .2)	(.4, .2, .4)	.285	.299	.535	.074
2	(.1, .4, .4, .1)	(.4, .3, .2, .1)	.528	*	.0078	*
3	(.1, .2, .2, .1, .2, .2)	(.05, .1, .15, .2, .2, .3)	.636	*	*	*

Table 1: System parameters for the simulation study, and resulting  $p$ -values of the Hotelling’s  $T^2$  test. Asterisks denote  $p$ -value  $< 10^{-15}$ .

In each simulation replication, the system was initialized with all servers simultaneously starting service of successive customers, each customer being randomly chosen from the server’s compatibility set. To let the system approach steady state, it was first run for 100,000 service completions as a warmup period. After warmup, the system was run for additional 1,000,000 service completions, and the fraction of services of customer type  $c_i$  by server  $s_j$  was recorded. This procedure was repeated 100 times, and each element of the

final estimated matrix  $\hat{r}$  is therefore a mean of 100 simulated fractions. The simulation was carried out using the R programming language ([www.r-project.org](http://www.r-project.org)).

For each model, the matrix  $r$  was analytically computed as described in [4], and the estimated matrices  $\hat{r}_p$ ,  $\hat{r}_{u_1}$ , and  $\hat{r}_{u_2}$  were computed by simulation. The resulting matrices are shown below. The entries of the estimated matrices are invariably very close to the theoretical  $r$  values, yet when comparing them using Hotelling's  $T^2$  test (see below), it turns out in most cases that they are different in a statistically significant manner; see the last three columns of Table 1. Interestingly, the matching rates in system 1 appear to be insensitive to the service time distribution. As a control for the veracity of our simulation, we simulated the system also under exponential service time distribution, and as expected, did not get any significant test results; see column 4 of Table 1.

Hotelling's  $T^2$  test is the multivariate generalization of the ubiquitous Student's  $t$  test. In each simulation replication, the non-zero entries of the empirical matching rate matrix  $\hat{r}$  (those corresponding to service compatibility) may be thought of as a realization of a random vector. The entries of this vector, however, are dependent, as they must sum to 1. The null hypothesis of the test is that the mean of this vector is the corresponding vector derived from the theoretical matching rate matrix  $r$ . The test's statistics  $T^2$  is a scaled sum of the squared deviations of the observed vectors from the hypothesized mean vector; under the null hypothesis, it possesses asymptotically an  $F$  distribution. To make the empirical covariance matrix of the observed (simulated) vectors invertible, the last entry of each vector was omitted. For more details on Hotelling's  $T^2$  test, see [24].

### System 1

$$r = \begin{pmatrix} 0 & .1 & .1 \\ .3 & 0 & .3 \\ .1 & .1 & 0 \end{pmatrix}, \quad \hat{r}_p = \begin{pmatrix} 0 & .09996 & .10006 \\ .29987 & 0 & .30013 \\ .1 & .09997 & 0 \end{pmatrix}$$

$$\hat{r}_{u_1} = \begin{pmatrix} 0 & .1 & .09996 \\ .30002 & 0 & .30003 \\ .10001 & .09998 & 0 \end{pmatrix}, \quad \hat{r}_{u_2} = \begin{pmatrix} 0 & .10002 & .10009 \\ .29999 & 0 & .29991 \\ .10002 & .09998 & 0 \end{pmatrix}$$

### System 2

$$r = \begin{pmatrix} .06443 & 0 & 0 & .03557 \\ .3356 & .06443 & 0 & 0 \\ 0 & .2356 & .1644 & 0 \\ 0 & 0 & .03557 & .06443 \end{pmatrix}, \quad \hat{r}_p = \begin{pmatrix} .06477 & 0 & 0 & .03524 \\ .33519 & .0648 & 0 & 0 \\ 0 & .23523 & .16478 & 0 \\ 0 & 0 & .03531 & .06468 \end{pmatrix}$$

$$\hat{r}_{u_1} = \begin{pmatrix} .06447 & 0 & 0 & .03553 \\ .33549 & .06447 & 0 & 0 \\ 0 & .23556 & .16447 & 0 \\ 0 & 0 & .03554 & .06446 \end{pmatrix}, \quad \hat{r}_{u_2} = \begin{pmatrix} .06465 & 0 & 0 & .03537 \\ .33535 & .06461 & 0 & 0 \\ 0 & .2354 & .16465 & 0 \\ 0 & 0 & .03535 & .06463 \end{pmatrix}$$

### System 3

$$\begin{aligned}
r &= \begin{pmatrix} .004584 & .009497 & 0 & 0 & 0 & .08592 \\ .03825 & .06356 & .09819 & 0 & 0 & 0 \\ 0 & .02694 & .04084 & .1322 & 0 & 0 \\ 0 & 0 & .01097 & .02815 & .06087 & 0 \\ 0 & 0 & 0 & .03963 & .06184 & .09854 \\ .007164 & 0 & 0 & 0 & .07729 & .1155 \end{pmatrix} \\
\hat{r}_p &= \begin{pmatrix} .00462 & .01039 & 0 & 0 & 0 & .08499 \\ .03853 & .06318 & .0983 & 0 & 0 & 0 \\ 0 & .02638 & .04062 & .13298 & 0 & 0 \\ 0 & 0 & .01098 & .02787 & .0612 & 0 \\ 0 & 0 & 0 & .03923 & .06149 & .09927 \\ .00681 & 0 & 0 & 0 & .07741 & .11575 \end{pmatrix} \\
\hat{r}_{u_1} &= \begin{pmatrix} .00465 & .00894 & 0 & 0 & 0 & .08645 \\ .03781 & .06386 & .09834 & 0 & 0 & 0 \\ 0 & .02718 & .04078 & .132 & 0 & 0 \\ 0 & 0 & .0109 & .02819 & .06092 & 0 \\ 0 & 0 & 0 & .03979 & .06198 & .09823 \\ .00754 & 0 & 0 & 0 & .07713 & .11531 \end{pmatrix} \\
\hat{r}_{u_2} &= \begin{pmatrix} .00476 & .00864 & 0 & 0 & 0 & .08661 \\ .0374 & .06409 & .09857 & 0 & 0 & 0 \\ 0 & .02727 & .04055 & .13213 & 0 & 0 \\ 0 & 0 & .01088 & .02802 & .0611 & 0 \\ 0 & 0 & 0 & .03985 & .06206 & .09811 \\ .00784 & 0 & 0 & 0 & .07684 & .11529 \end{pmatrix}
\end{aligned}$$

A similar phenomenon occurs with the steady-state distribution of the *server span*  $Y_J(t) - Y_1(t)$ , which is the distance between the leftmost and rightmost servers along the stream of customers (note that the minimal value of the server span is  $J - 1$ ). Figure 7 shows the distribution of the server span for the three systems under the same four service time distributions, as estimated from simulation. Clearly, the distribution in each system is sensitive to the service time distribution.

From [2], the steady-state distribution of the server permutations in the exponential case is given by

$$\pi_R(S_1, \dots, S_J) = B^s \prod_{\ell=1}^{J-1} (\beta_{\{S_1, \dots, S_\ell\}} - \alpha_{\mathcal{U}\{S_1, \dots, S_\ell\}})^{-1},$$

where  $B^s$  is a normalizing factor. This distribution was estimated by simulation also for the non-exponential cases, and the results for systems 1 and 2 are shown in Tables 2 and 3 (the results for system 3 are omitted due to the size of the table —  $6! = 720$  rows). The deviations of the estimated values from the theoretical ones are small, but statistically significant: when using again Hotelling's  $T^2$  test, the  $p$ -values in all 6 cases ( $2 \text{ systems} \times 3 \text{ distributions}$ ) was  $< 10^{-15}$ . In contrast, the  $p$ -values for systems 1 and 2 under simulated exponential service times were 0.372 and 0.443, respectively. Thus, the steady-state distribution of the server permutations is also sensitive to the service time distribution.

## References

- [1] Adan, I., Boon, M., and Weiss, G., (2014) A design heuristic for skill based parallel service systems. Preprint

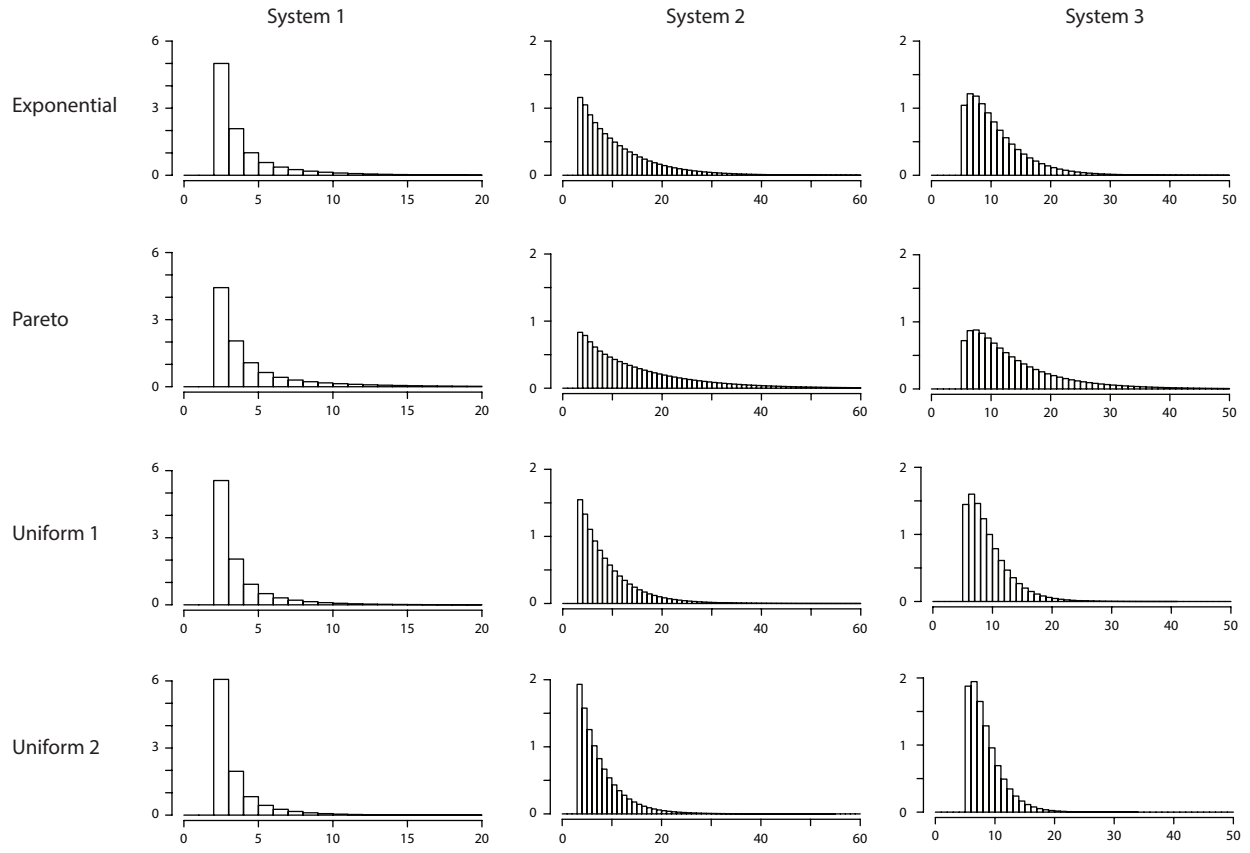


Figure 7: Server span distribution for the three systems under four service time distributions. Counts on the vertical axes are  $\times 10^7$ .



permutation	theoretical (exponential)	Pareto	Uniform 1	Uniform 2
1-2-3	.1	.1018	.0973	.0912
1-3-2	.2	.1996	.2006	.2010
2-1-3	.2	.1988	.2021	.2078
2-3-1	.2	.1988	.2021	.2078
3-1-2	.2	.1993	.2006	.2010
3-2-1	.1	.1017	.0974	.0912

Table 2: Steady-state distribution of server permutations, system 1.

- [2] Adan, I., Busic, A., Mairesse, J., Weiss, G. (2015). Reversibility and further properties of FCFS infinite bipartite matching. *arXiv:1507.05939*.
- [3] Adan, I., Foss, S., Shneer, S., Weiss, G. (2015). Local stability in a transient Markov chain. *arXiv:1511.06094*.
- [4] Adan, I., Weiss, G. (2011) Exact FCFS matching rates for two infinite multi-type sequences. *Operations Research*, **60** 475–489.
- [5] Adan, I., Weiss, G. (2014) A queue with skill based service under FCFS-ALIS: steady state, overloaded system, and behavior under abandonments. *Stochastic Systems*, **4**(1):250-299.
- [6] Ahuja, R.K., Magnanti, T.L., Orlin, J.B. (1993). *Network Flows: Theory, Algorithms, and Applications* Prentice Hall
- [7] Armony, M., Ward, A.R. (2010). Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research*, **58**(3), 624-637.
- [8] Armony, M., Ward, A.R. (2013). Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research*, **61**(1), 228-243.
- [9] Bell, S.L., Williams, R.J. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *The Annals of Applied Probability*, **11**(3), 608-649.
- [10] Bramson, M. (2008). *Stability of Queueing Networks*. Springer, Berlin, Heidelberg.
- [11] Caldentey, R., Kaplan, E.H., Weiss, G. (2009) FCFS infinite bipartite matching of servers and customers. *Advances in Applied Probability*, **41** 695–730.
- [12] Dai, J.G. (1995) On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability*, **5**(1), 49-77.
- [13] Dai, J.G., Lin, W. (2005). Maximum pressure policies in stochastic processing networks. *Operations Research*, **53**(2), 197-218.
- [14] Dai, J.G., Weiss, G. (1996). Stability and instability of fluid models for reentrant lines. *Mathematics of Operations Research*, **21**(1):115-134.
- [15] Foss, S., Chernova, N. (1998). On the stability of a partially accessible multistation queue with state-dependent routing. *Queueing Systems*, **29**(1), 55-73.
- [16] Gans, N., Koole, G., Mandelbaum, A. (2003). Telephone call centers: tutorial, review, and research prospects. *Manufacturing Service Operations Management*, **5**(2) 79-141.
- [17] Ghamami, S., Ward, A.R. (2013). Dynamic scheduling of a two-server parallel server system with complete resource pooling and reneging in heavy traffic: Asymptotic optimality of a two-threshold policy. *Mathematics of Operations Research*, **38**(4):761-824.

permutation	theoretical (exponential)	Pareto	Uniform 1	Uniform 2
1-2-3-4	.0232	.0289	.0194	.0169
1-2-4-3	.0077	.0089	.0068	.0061
1-3-2-4	.0116	.0133	.0106	.0103
1-3-4-2	.0023	.0026	.0022	.0022
1-4-2-3	.0058	.0083	.0045	.0038
1-4-3-2	.0035	.005	.0027	.0024
2-1-3-4	.0310	.0286	.0319	.0323
2-1-4-3	.0103	.0092	.0110	.0116
2-3-1-4	.0929	.0933	.0888	.0839
2-3-4-1	.0929	.1042	.0849	.0792
2-4-1-3	.0077	.0064	.0088	.0096
2-4-3-1	.0232	.0215	.0243	.0250
3-1-2-4	.0232	.0201	.0251	.0257
3-1-4-2	.0046	.0039	.0052	.0054
3-2-1-4	.1394	.1319	.1451	.1499
3-2-4-1	.1394	.1426	.1416	.1456
3-4-1-2	.0139	.0120	.0150	.0153
3-4-2-1	.0697	.0680	.0695	.0690
4-1-2-3	.0232	.0231	.0223	.0209
4-1-3-2	.0139	.0137	.0137	.0137
4-2-1-3	.0232	.0199	.0260	.0276
4-2-3-1	.0697	.0686	.0692	.0671
4-3-1-2	.0279	.0248	.0301	.0313
4-3-2-1	.1394	.1411	.1410	.1451

Table 3: Steady-state distribution of server permutations, system 2.

- [18] Green, L. (1985) A queueing system with general-use and limited-use servers, *Operations Research*, **33**:162–182.
- [19] Gurvich, I., Whitt, W. (2009). Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research*, **34**(2), 363-396.
- [20] Gurvich, I., Whitt, W. (2010). Service-level differentiation in many-server service system via queue-ratio routing. *Operations Research*, **58**(2), 316-328.
- [21] Harchol-Balter, M., Crovella, M.E., Murta, C.D. (1999). On choosing a task assignment policy for a distributed server system. *Journal of Parallel and Distributed Computing*, **59**(2), 204-228.
- [22] Harrison, J.M., Lopez, M.J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing systems*, **33**(4), 339-368.
- [23] Harrison, J.M., Zeevi, A. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management*, **7**(1), 20-36.
- [24] Krzanowski, W.J. (2000), *Principles of Multivariate Analysis*, Oxford University Press.
- [25] Meyn, S.P., Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. Springer.
- [26] Rubino, M., Ata, B. (2009). Dynamic control of a make-to-order, parallel-server system with cancellations. *Operations Research*, **57**(1), 94-108.
- [27] Squillante, M.S., Xia, C.H., Yao, D.D., Zhang, L. (2001). Threshold-based priority policies for parallel-server systems with affinity scheduling. In American Control Conference, 2001. *Proceedings of the 2001: 2992-2999*, IEEE.

- [28] Talreja, R., Whitt, W. (2008) Fluid models for overloaded multi-class many-service queueing systems with FCFS routing. *Management Science*, **54**, 1513–1527.
- [29] Tezcan, T., Dai, J.G. (2010). Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research*, **58**(1):94-110.
- [30] Vanderbei, R.J. (2001). *Linear Programming, Foundations and Extensions*. Second Edition, International Series in Operations Research and Management Science, **37**.
- [31] Veeger, C.P.L., Etman, F.P. and Rooda, V. (2008) Generating cycl time-throughput-product mix surfaces using effective process time based aggregate modeling. *Proc. 13th ASIM Conf., Berlin*, 519-529.
- [32] Visschers, J., Adan, I., G. Weiss, G. (2012) A product form solution to a system with multi-type customers and multi-type servers. *Queueing Systems*, **70**, 269–298.
- [33] Wallace, R.B., Whitt, W. (2005). A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management*, **7**(4), 276-294.
- [34] Williams, R.J. (2000). On dynamic scheduling of a parallel server system with complete resource pooling. *Fields Institute Communications*, **28**, 49-71.